Statistics in Biosciences
https://doi.org/10.1007/s12561-023-09400-x

AI-Powered Bayesian Statistics in Biomedicine

Qiwei Li[1]

Received: 10 September 2023 / Revised: 10 September 2023 / Accepted: 5 October 2023

Abstract
Statistics and artificial intelligence (AI) are distinct yet closely interconnected disciplines, each characterized by its own historical roots and methodological approaches. This paper explores their collaborative potential, seeking to answer a pivotal question: How can statistics and AI collaborate to extract valuable insights from complex data? Within this context, we present three compelling case studies that showcase the harmonious integration of statistics and AI for the analysis of high-resolution pathology images, an emerging type of medical image that provides rich cellular-level information and serves as the gold standard for cancer diagnosis. Furthermore, recent advancements in spatial transcriptomics, which typically yield paired digital pathology images from the same tissue sample, introduce a new dimension to pathology images. This evolving landscape extends the horizons of the proposed AI-statistics framework, holding a promise of propelling biomedical research into new territories and delivering breakthroughs in our understanding of complex diseases.

Keywords Artificial intelligence · Bayesian statistics · Spatial analysis · Shape analysis · Pathology image

1 Introduction

Statistics is not a new discipline of science or technology. The term *statistics* was introduced by an Italian writer, Girolamo Ghilini, in 1589 [1, 2]. As a branch of mathematics, statistics began evolving about three centuries ago in response to the novel needs during the First Industrial Revolution. Statistical inferences are made under the framework of probability, another branch of mathematics dealing with random phenomena analysis, dating back to earlier times. For instance, Bayesian statistics, one of the pivotal branches of statistics, is named after Thomas Bayes, who

✉ Qiwei Li
qiwei.li@utdallas.edu

[1] Department of Mathematical Sciences, The University of Texas at Dallas, 800 W Campbell Rd, Richardson TX, 75035, USA

Published online: 26 October 2023

was the first to use probability inductively and formulate a specific case of Bayes' theorem, a fundamental theorem in probability. The foundations of modern statistics were further fortified by Karl Pearson, who also established the world's first university statistics department at University College London in 1911. His seminal work in the early 20th century underpins many of the classical statistical methods that are in common use today, such as correlation coefficient, *p*-value, Chi-squared test, principal component analysis (PCA), etc.

In contrast, artificial intelligence (AI) is a new academic discipline. The term *machine learning* (ML) was coined in 1959 by Arthur Samuel [3], an IBM employee and pioneer in computer gaming and AI. ML has been recognized as an integral component of AI, as AI studies how machines can imitate the intelligence or behavioral pattern of humans or any other living entity, and ML refers to any AI technique by which a machine can learn from data without using a complex set of different rules. In the late 2000s, deep learning, a type of ML technique inspired by the human brain's network of neurons, started to outperform other methods in ML competitions. The breakthroughs of deep learning have motivated people to rethink how to integrate information, analyze big data, and improve decision-making. AI is revolutionizing various industries because of its huge impact on every walk of life. While AI, ML, and deep learning are technically different, the three terms will be used interchangeably throughout this paper.

The major difference between statistics and AI are three-fold. First and foremost, statistics is a mathematical body of science based on probability theorems, seeking to objectively explain the data of nature in a reproducible way, while AI is an engineering that applies natural science or mathematics to solve real-world problems. Machine learning also has intimate ties to optimization, which has been widely used in engineering. For instance, many ML problems can be formulated as minimization of some loss function on training data. Secondly, although both statistics and ML is the mathematical study of data, their overarching objectives diverge significantly. Statisticians typically focus on building a generalized model to fit all kinds of data and studying the goodness of fit, while ML engineers aim to discover complex patterns in data. Lastly, they are two opposite approaches [4]. Statistical methods are typically top-down approaches. We assume the model that generates the observed data is known and the probabilistic dependency between the model and data build upon predetermined equations with simple assumption. ML methods, in contrast, are bottom-up approaches. No particular predetermined model or equation is assumed, but one begins with the data and an algorithm develops a method to perform better in a specific supervised or unsupervised learning task (e.g., classification or clustering).

We are living in the era of big data. Which one is better for modern big data analytics, statistics or AI? Although there are some debates between the two options, the answer depends on the study goal. ML is probably the best pick to achieve peak performance in a supervised learning task where well-trained data are available. Conversely, when the objective is to establish relationships among variables or extract meaningful interpretations from data with a small sample size, statistical models rise to the fore. This is because statistics and AI have opposite strengths and weaknesses. AI shines at discovering complex patterns from data, but lacks

interpretability and reproducibility to some extent. In contrast, statistical inference delivers clear and interpretable results but tends to rely on assumptions that could oversimplify complex data structures. In many instances, these two choices need not stand in opposition; they can harmoniously coexist. This is especially true when the interpretation of noisy and complex data is in great need to advance new scientific discovery. Indeed, statistics and AI stand as closely related fields. As suggested by some leading statisticians, we need a term such as *data science* or *statistical learning* as a placeholder to call the overall area [5, 6].

In this review paper, we introduce three examples of combining statistics and AI to seamlessly analyze pathology images, a type of high-resolution medical image that captures histological details and serves as the golden standard in cancer diagnosis and prognosis. A tumor pathology image, also known as a whole slide image (WSI), harbors a large amount of information at the cellular level, such as interactions between tumor cells and the surrounding micro-environment. This routine clinical procedure produces massive digital pathology images on a daily basis. However, the exhaustive and time-consuming process of manual pathological examination, reliant on human expertise, has, until now, imposed limitations on the systematic and comprehensive exploration of these high-resolution images. Moreover, recent breakthroughs in spatial transcriptomics (ST) have enabled the molecular and spatial characterizations of single cells. As this cutting-edge technique typically yields paired pathology images from the same tissue sample, we can view the spatial molecular profiling data as a new dimension to the pathology image. Because of the complexity of the data and the emerging need for data interpretation, neither AI-based nor statistical methods can face the great challenges alone in this field.

To this end, we proposed a unified AI-statistics framework, as depicted in Fig. 1(a) to analyze pathology images, leveraging the strengths of AI and statistics. Our idea is first to tailor the deep learning method to denoise complex imaging data in a specific task and abstract its simple representation. Subsequently, we formulate statistical models to fit the AI-processed data and use the estimated model parameters to make interpretations and perform further association studies with other datasets of interest. Within this framework, we express a strong preference for Bayesian statistics. On one hand, Bayesian inference has shown great success in analyzing biomedical data [7–14]. On the other hand, it is able to make more inferences and utilize the existing prior information, especially in high-dimensional settings where large samples are unachievable. Although the proposed project is rooted in analyzing pathology images, we can apply the proposed AI-statistics framework to analyze complex data in a broad range of disciplines.

The remainder of the paper is organized as follows: Sect. 2 illustrates an example of quantifying cell–cell interaction from pathology images using a model-based approach [13]. Section 3 presents a case study that applies a novel statistical shape analysis method to characterize tumor boundaries [15]. In Sect. 4, we demonstrate that fully exploiting the morphological features present in pathology images and the molecular features measured by ST can enhance the accuracy and interpretability of spatial domain identification [16], a central challenge for ST data analysis. We follow the order below to illustrate each of the three examples: (1) a high-level summary of the project with research goals; (2) a summary of the AI-process data; (3) a

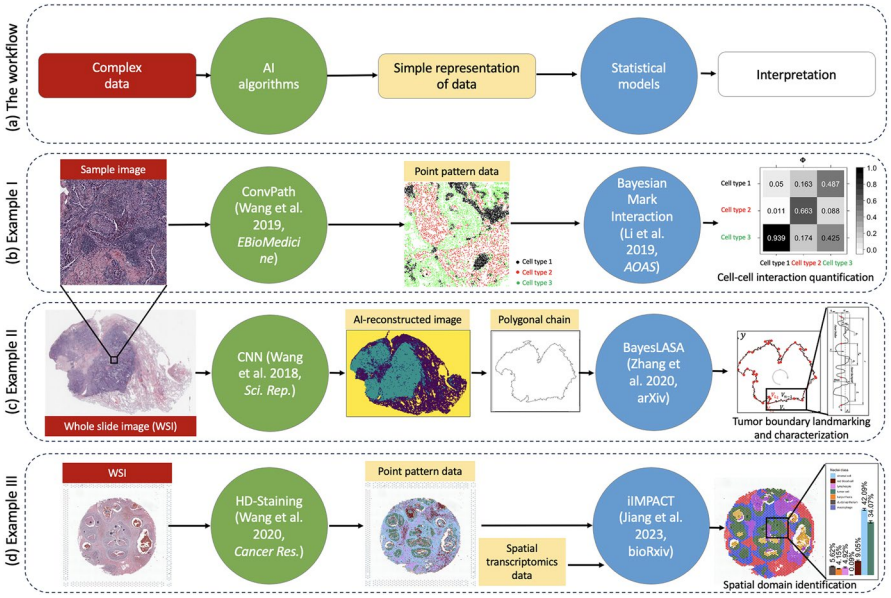

Fig. 1 **a** The workflow of combining statistics and AI to analyze complex data. **b** The illustration of Example I, applying the workflow to quantify cell–cell interaction. **c** The illustration of Example II, applying the workflow to characterize tumor boundary. **d** The illustration of Example III, applying the workflow to identify histology-based spatial domains

discussion of current methods and challenges; (4) a brief presentation of the Bayesian model and its key highlights; and (5) an overview of the outcomes. Our intention is to provide a snapshot of the depth and breadth of our AI-statistics framework. We encourage interested readers to read the original papers for a more profound exploration of our research endeavors. Section 5 concludes the paper with some remarks.

2 Example I: Quantifying Cell–Cell Interaction

The primary goal of this study is to quantify the interactions between different types of cells within tumor regions of a WSI. To extract the cell information, we first developed a deep learning-based pipeline to identify individual cells and classify their types into different categories. Then, we consider such an AI-reconstructed image as multi-type point pattern data. A novel model-based approach through a Bayesian framework, was proposed to analyze spatial correlations of cell types conditional on their locations. Figure 1b illustrates the workflow of this project.

2.1 AI-Processed Data

We first used a convolutional neural network (CNN)-based method, ConvPath [17], to locate each cell and predict its cell types (i.e., lymphocyte, stromal, and tumor

cells) from a 5000×5000 pixel window, namely, a sample image, in the tumor region of a given WSI. This study includes 1585 sample images from 188 lung cancer patients' WSI (with a median size of $24{,}244 \times 19{,}261$ pixels) in the National Lung Screening Trial (NLST) study. As a result, each sample image was abstracted into a spatial map of marked points, where each point indexed by $i = 1, \ldots, n$ refers to a cell at location $(x_i, y_i) \in \mathbb{R}^2$ and its qualitative mark denoted by $z_i \in \{1, \ldots, Q\}$. In spatial point pattern analysis, such data are considered as multi-type point pattern data. Here, the number of cells per sample image, denoted by n, ranges from 2, 876 to 26, 463, and the number of cell types $Q = 3$.

2.2 Current Methods and Challenges

The study of interactions between qualitative or quantitative marks, which results in the spatial correlation of marks, has been a primary focus in spatial statistics. Illian et al. [18] discussed in detail a large variety of numerical, functional, and second-order summary characteristics, which can be used to describe the spatial dependency between different types of points in a planar region. However, model-based analysis, which may sharpen inferences about the spatial correlation, is lagging. Using the same dataset described above, Li et al. [19] and Li et al. [20] modified the Potts model, a model of interacting spins on a lattice, to indirectly quantify the cell–cell interaction. However, the main issues are that these approximate methods relies on selecting an *ad hoc* lattice and do not directly model the spatial correlation of cell types at the cellular level.

2.3 Methods

To this end, we developed a novel Bayesian mark interaction model to study the mark formulation at a finite known set of points through a Bayesian framework [13]. The key idea is to propose a well-defined energy function that minimizes the overall energy of the cell–cell interaction network,

$$\begin{aligned} H(\boldsymbol{z}|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda) = & \sum_{q=1}^{Q} \omega_q \sum_{i=1}^{n} I(z_i = q) \\ & + \sum_{q=1}^{Q} \sum_{q'=1}^{Q} \theta_{qq'} \sum_{i<i'} I(z_i = q, z_{i'} = q') \exp(-\lambda d_{ii'}) I(d_{ii'} < c), \end{aligned} \tag{1}$$

where $\boldsymbol{z} = [z_i \in \{1, \ldots, Q\}]_{n\times 1}$ records n cells' types, $d_{ii'} \in \mathbb{R}^+$ is the distance between cell i and i', $\lambda \in \mathbb{R}^+$ is the parameter that mimics the exponential decay of pairwise interaction energy with distance, $I(\cdot)$ denotes the indicator function, and $c \in \mathbb{R}^+$ is the desired threshold meaning that any cell can only interact with its nearby cells within a certain range. Parameters $\boldsymbol{\omega} = [\omega_q \in \mathbb{R}]_{Q\times 1}$ and $\boldsymbol{\Theta} = [\theta_{qq'} \in \mathbb{R}]_{Q\times Q}$ are the first and second-order intensities, indicating the enrichment of different types of cells and the spatial correlation among different types of cells, respectively. According to the Hammersley-Clifford theorem

[21], such a local-defined energy leads to the data likelihood $f(z_1, \ldots, z_n|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda) = H(z|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda)/\sum_{z'} H(z'|\boldsymbol{\omega}, \boldsymbol{\Theta}, \lambda)$. Note that an exact calculation of the denominator needs to sum over the entire space of $\boldsymbol{z}$ consisting of Q^n states. To overcome this challenge, we employed the double Metropolis-Hastings (DMH) algorithm [22] to make inference on the model parameters $\boldsymbol{\omega}$, $\boldsymbol{\Theta}$, and λ. We showed that Equation (1) can be directly linked to the conditional distribution as $\Pr(z_i = q|z_{i'} = q') \propto \exp(-\theta_{qq'})$ in the simplest scenario, implying the spatial correlations among marks can be easily interpreted by the probability matrix $\boldsymbol{\Phi} = [\phi_{qq'} \in [0, 1]]_{Q\times Q}$ with $\phi_{qq'} = \exp\left(-\theta_{qq'}\right)/\sum_{q=1}^{Q} \exp\left(-\theta_{qq'}\right)$.

2.4 Results

The real data analysis shows that the spatial correlation between tumor and stromal cells is associated with patient prognosis (p-value=0.007) through a Cox proportional hazards model [23]. Although the morphological features of stroma in tumor regions have been discovered to be associated with patient survival [24], there is no strong quantitative evidence to support this, due to a lack of rigorous statistical methodology. The Bayesian mark interaction model delivers a new perspective for understanding how marks (i.e. cell types in AI-reconstructed pathology images) formulate. This estimated spatial corrections in terms of $\boldsymbol{\Phi}$ could be translated into real clinical tools at low cost because it is based only on tumor pathology slides, which are available in standard clinical care.

3 Example II: Characterizing Tumor Boundary

This study aims to characterize the heterogeneous boundary roughness of tumor regions in a WSI. An automated tumor region recognition system based on deep CNN was developed. Then, we extracted the tumor boundary from the AI-reconstructed image. Considering the tumor boundary, a sequence of pixel points, as a simple closed polygonal chain (SCPC), we developed a Bayesian landmark-based shape analysis to estimate the number and locations of its landmarks, which helps to summarize the heterogeneous tumor boundary roughness. Figure 1(c) illustrates this study's workflow.

3.1 AI-Processed Data

We collected 246 WSI from 143 lung cancer patients in the NLST study. We first developed a CNN-based method [25] to perform tumor segmentation for each WSI. This approach classifies each 300×300 pixels image patch in a WSI into three categories: normal, tumor, or background. The median size of the resulting three-class AI-reconstructed images is 1011×806 pixels. Then, we enlarged each image three times (to avoid single-pixel boundary lines or singularities) and used the `R` package `SAFARI` [26] to extract the largest connected tumor region and its boundary from each image. The tumor boundary was abstracted into a sequence of m discretization points denoted

by $P = \{V_1, \ldots, V_m\}$, forming an SCPC. The coordinates of V_i is $(x_i, y_i) \in \mathbb{R}^2$. In this study, the number of SCPC vertices ranges from $n - 1 = 360$ to 15, 931 across all the 246 WSI, with a median of 3836.

3.2 Current Methods and Challenges

Traditional shape features that characterize an object's boundary roughness are based on radial lengths, which have improved clinical diagnosis [27–30] and prognosis [31, 32]. However, they have been recently proven to be no longer suitable for high-resolution pathology images at the cellular level [26], which exhibit substantial heterogeneity. To overcome this challenge, our idea is to identify a set of landmarks to partition the entire boundary into pieces based on roughness. Landmark identification has been a primary focus in shape analysis. Recently, a Bayesian model has been proposed under the elastic curve paradigm [33]. Functional data are infinite-dimensional, which raise computational concern in analyzing complex tumor shapes in high-resolution pathology images.

3.3 Methods

To characterize detailed and heterogeneous tumor boundary structures, we developed a novel Bayesian model, namely BayesLASA, to partition the entire boundary by a set of landmarks based on both the global geometry and local roughness [15]. To begin with, we use a latent binary vector $\boldsymbol{\gamma} = [\gamma_i \in \{0, 1\}]_{m\times 1}$ to indicate which vertices are landmarks, with $\gamma_i = 1$ if vertex i is a landmark. Those landmarks constitute another SCPC named the landmark chain $P^{(\gamma)} = \{V_{L_1}, \ldots, V_{L_K}\}$, where we use L_k to denote the location of landmark k and $K = \sum_{i=1}^{m} \gamma_i$ is the number of landmarks. Since $\boldsymbol{\gamma}$ is independent of the vertex locations, it is naturally invariant to rotation, scaling, translation, and other shape-preserving transformations. From another point of view, those non-landmark vertices can be assigned into pieces bounded by two adjacent landmarks. Thus we use $\boldsymbol{\xi} = [\xi_i \in \{1, \ldots, K\}]_{m\times 1}$ to reparameterize $\boldsymbol{\gamma}$, where $\xi_i = k$ if vertex i is between landmarks V_{L_k} and $V_{L_{k+1}}$. The objective is to find the landmark chain $P^{(\gamma)}$ via inferring $\boldsymbol{\gamma}$ or $\boldsymbol{\xi}$, which are identical. To enable the model to identify landmarks based on local roughness, we assume the vertex-wise deviation between P and $P^{(\gamma)}$ is from a mixture zero-mean stationary Gaussian Process. In particular, let d_i denote the shortest distance between V_i and the line segment between $V_{L_{\xi_i}}$ and $V_{L_{\xi_i+1}}$ in $P^{(\gamma)}$, then we have,

$$d_1, \ldots, d_m | \boldsymbol{\xi}, \sigma_1^2, \ldots, \sigma_K^2 \sim \prod_{k=1}^{K} \mathrm{N}((d_{L_k+1}, \ldots, d_{L_{(k+1)}-1})^\top; \mathbf{0}, \sigma_k^2 \boldsymbol{\Sigma}_k), \tag{2}$$

where σ_k^2 is a piecewise scaling factor, indicating the average deviation between P and $P^{(\gamma)}$ in piece k and $\boldsymbol{\Sigma}_k$ is a covariance function of the pairwise distances. Note that we integrate out σ_k^2's so that the number of landmarks K can be automatically quantified through $\boldsymbol{\gamma}$. For the sake of simplicity, we chose the white noise kernel, where $\boldsymbol{\Sigma}_k$ is an identity matrix. The identified landmarks, which approximately reconstruct the tumor shape, partition the whole boundary into mutually exclusive

pieces. Summary statistics of the piecewise roughness measurements can then be used to characterize the heterogeneity of boundary roughness.

3.4 Results

The real data result shows that the heterogeneity (in terms of skewness and kurtosis) of tumor boundary roughness is significantly associated with patient prognosis (p-value < 0.001) through a Cox proportional hazards model [23]. These results are consistent with the biology literature in that high spatial heterogeneity is a pivotal feature of cancer at both the cellular and histological levels resulting from the distinct patterns of different cancer cell subpopulations in terms of dysregulation of proliferation, mobility, and metabolism pathways [34, 35]. Analyzing the same datasets, Moon et al. [36] developed a functional representation of tumor topological structure, pairing those topological features with the surrounding environment using the persistent homology. The results show that the topological features also predict survival prognosis.

4 Example III: Identifying Spatial Domains

Examples I and II focus on analyzing AI-reconstructed pathology images only, while Example III introduces an integrative model to enhance ST clustering analysis, an essential task in this emerging field, by fully exploiting the morphological features in pathology images. Firstly, we developed a mask regional CNN (Mask R-CNN)-based algorithm to identify all individual cells in a WSI and classify their types. We then summarized this large-scale multi-type point pattern data at the same spatial resolution as the paired ST data. Lastly, a Bayesian finite mixture model (FMM) was proposed to integrate these two modalities of a tissue sample and partition all spots into mutually exclusive clusters, namely spatial domains.

4.1 AI-Processed Data

ST captures RNA molecules *via* spatially arrayed barcoded probes, namely spots, which cover a group of cells and are arrayed on a two-dimensional grid. In general, the molecular profile of ST data can be represented by $\boldsymbol{Y} = [y_{ij} \in \mathbb{N}]_{n\times p}$ with y_{ij} is the read count for gene j measured at spot i. Let $(x_i, y_i) \in \mathbb{R}^2$ be the coordinates of spot i. Since ST spots are on a lattice grid, a convenient way to define the geospatial profile is *via* a binary adjacent matrix $\boldsymbol{G} = [g_{ii'} \in \{0, 1\}]_{n\times n}$ with $g_{ii'} = 1$ if spot i are i' are neighbors in the grid and $g_{ii'} = 0$ otherwise. To construct the image profile, we applied a nuclei segmentation and identification algorithm, the histology-based digital (HD)-staining model [37], to extract the location and type of each cell from the paired WSI. The HD-staining model was trained using the pathology images in the NLST study and implemented by the Mask R-CNN architecture. To match the molecular profile at the spot level, we count cells with different types within each spot and create a cell abundance table, denoted by $\boldsymbol{V} = [v_{iq} \in \mathbb{N}]_{n\times Q}$, where each entry v_{iq} is the number of cells

with type q observed at spot i. Our first case study is to apply our method to a human breast cancer dataset, which includes $n = 2,518$ spots and $p = 17{,}651$ genes. There are 156, 235 cells identified by HD-staining in $Q = 7$ categories (i.e., macrophage, ductal epithelium, karyorrhexis, tumor, lymphocyte, red blood, and stromal cells).

4.2 Current Methods and Challenges

A central challenge for ST data analysis is to define clinically or biologically meaningful spatial domains by partitioning regions with similar molecular and/or histological characteristics, because the spatial domain identification serves as the foundation for several important downstream analyses [38, 39]. However, current state-of-the-art methods typically focus on achieving this goal solely by analyzing molecular profiles [40, 41]. There are several recently developed methods integrating spatial information and various features extracted from the histology image into the clustering analysis of ST data [42–45]. However, those image features, such as RGB color values, do not explicitly reveal detailed morphological information, and therefore, fail to provide biologically relevant insights. Different from molecular information, pathology images characterize cellular structures and tissue microenvironment, which have been proven valuable in clinical diagnosis and prognosis [37, 46]. Thus, integrating molecular profiles and AI-reconstructed pathology images could enhance the spatial domain identification.

4.3 Method

A Bayesian FMM, namely iIMPACT, was developed to tackle this problem. Generally, an FMM generates random variables from a weighted sum of K independent distributions that belong to the same parametric family. Since there are two modalities $\boldsymbol{Y}$ and $\boldsymbol{V}$, we decomposed the mixture component into two sub-components,

$$\tilde{\boldsymbol{y}}_i, \boldsymbol{v}_i | z_i = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\omega}_k \sim \mathrm{N}(\tilde{\boldsymbol{y}}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\mathrm{Multi}(\boldsymbol{v}_i; m_i, \boldsymbol{\omega}_k)^w, \tag{3}$$

where $\boldsymbol{z} = [z_i \in \{1, \ldots, K\}]_{n\times 1}$ denotes the latent variables specifying the spatial domain with $z_i = k$ indicating spot i belongs to spatial domain k. The multivariate normal sub-component models the low-dimensional gene expression $\tilde{\boldsymbol{y}}_i$ at spot i. In particular, we performed PCA to obtain $\tilde{\boldsymbol{y}}_i$ to avoid using complex FNNs with feature selection based on cumbersome multivariate Poisson distribution. This modeling approach is similar in spirit to the recent proposal by Zhao et al. [44]. The multinational sub-component models the number of cells with different types, where $m_i = \sum_{q=1}^{Q} v_{iq}$ is the total number of cells observed within spot i and $\boldsymbol{\omega}_k = [\omega_{kq} \in [0, 1]]_{Q\times 1}$, defined on a Q-dimensional simplex, representing the underlying relative abundance of cell types in spatial domain k, which can be used to interpret or define the identified spatial domains. To utilize the geospatial profile, iIMPACT employs a Markov random field (MRF) prior on the domain indicator $\boldsymbol{z}$, encouraging neighboring spots to be clustered into the same spatial domain. Of particular note is that $w \in [0, 1]$ controls the image profile's contribution to the clustering process, with respect to that of the molecular profile. Uncertainty quantification

is one advantage of the proposed Bayesian FFN. We define the spot as the boundary spot if the marginal probabilities $\Pr(z_i = k|\cdot)$ is small for any k, and the resulting connected area as the interactive zone.

4.4 Results

Compared with alternative methods, we found that iIMPACT achieved the highest consistency with the manual annotation, with an adjusted Rand index of 0.634. Moreover, iIMPACT is able to define each spatial domain k through its underlying relative abundance of cell types $\boldsymbol{\omega}_k$. In contrast, other methods currently lack the ability to effectively integrate cell type information and interpret the identified domains in a biologically meaningful way. We also demonstrated iIMPACT's superiority in a human prostate case study and a human ovarian case study, confirmed by ground truth biological knowledge. These findings underscore the accuracy and interpretability of iIMPACT as a new ST clustering approach, providing valuable insights into the cellular spatial organization.

5 Conclusion

Recent advancements in deep learning have enabled us to identify and classify individual cells or regions from digital pathology images on a large scale. This breakthrough paves the way for clarifying the many roles of cell–cell interaction, tumor shapes, and molecular features from these complex and rich data. Furthermore, it creates a unique opportunity for statisticians to foster statistical spatial and shape analysis rooted in model-driven research.

In this paper, we embark on a journey through three illustrative examples that showcase how the marriage between statistics and AI leads to more explainable and predictable paths from raw pathology images to conclusions. Example 1 concerns the spatial modeling of AI-reconstructed pathology images. The randomly distributed cells can be considered from a marked point process. A novel Bayesian model for characterizing spatial correlations in a multi-type spatial point pattern is presented. Example 2 concerns the statistical shape analysis. From the identified tumor regions in an AI-reconstructed pathology image, the tumor boundary is considered an SCPC. A novel Bayesian model is delivered to identify landmark points of the SCPC to provide descriptive statistics and characterize tumor boundary roughness. These two novel methodologies offer a unique perspective for comprehending the roles of cell–cell interactions and tumor growth patterns in the context of cancer progression. Example 3 concerns the integrative modeling of the emerging ST data, which comprehensively characterizes the molecular and morphological contexts at a high spatial resolution. In summary, these three examples collectively demonstrate how biomedical research can benefit from both statistics and AI.

Nowadays, statistics relies more on human analyses with computer aids, while AI relies more on computer algorithms with aids from humans. Nevertheless, expanding the statistics concourse at each milestone provides new avenues for AI

and creates new insides in statistics. This paper incubates the findings initiated from either side of statistics or AI and benefits the other. We envision that the proposed framework uniting statistics and AI in the analysis of complex data, will find applications across a myriad of disciplines, catalyzing innovation and insight.

Funding This study was supported by National Science Foundation (Grant Nos. 2113674, 2210912) and National Institutes of Health (Grant No. 1R01GM141519-01).

References

1. Ostasiewicz W (2014) The emergence of statistical science. Śląski Przegląd Statystyczny 18(12):75–82
2. Bruneau Q (2022) States and the masters of capital: sovereign lending, old and new. Columbia University Press, New York
3. Samuel AL (2000) Some studies in machine learning using the game of checkers. IBM J Res Dev 44(1.2):206–226
4. Ley C, Martin RK, Pareek A, Groll A, Seil R, Tischer T (2022) Machine learning and conventional statistics: making sense of the differences. Knee Surg Sports Traumatol Arthrosc 30(3):753–757
5. Hastie T, Tibshirani R, Friedman JH, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, vol 2. Springer, New York
6. Jordan MI, Mitchell TM (2015) Machine learning: trends, perspectives, and prospects. Science 349(6245):255–260
7. Li F, Zhang NR (2010) Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J Am Stat Assoc 105(491):1202–1214
8. Stingo FC, Vannucci M (2011) Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics 27(4):495–501
9. Stingo FC, Chen YA, Tadesse MG, Vannucci M (2011) Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. Ann Appl Stat 5(3):1978–2002
10. Do K-A, Qin ZS, Vannucci M (2013) Advances in statistical bioinformatics: models and integrative inference for high-throughput data. Cambridge University Press, Cambridge
11. Cassese A, Guindani M, Tadesse MG, Falciani F, Vannucci M (2014) A hierarchical Bayesian model for inference of copy number variants and their association to gene expression. Ann Appl Stat 8(1):148
12. Cassese A, Guindani M, Vannucci M (2016) iBATCGH: Integrative Bayesian analysis of transcriptomic and CGH data. In: Statistical analysis for high-dimensional data. Springer, Cham, pp 105–123
13. Li Q, Wang X, Liang F, Xiao G (2019) A Bayesian mark interaction model for analysis of tumor pathology images. Ann Appl Stat 13(3):1708
14. Li Q, Jiang S, Koh AY, Xiao G, Zhan X (2019) Bayesian modeling of microbiome data for differential abundance analysis. arXiv preprint. arXiv:1902.08741
15. Zhang C, Xiao G, Moon C, Chen M, Li Q (2020) Bayesian landmark-based shape analysis of tumor pathology images. arXiv preprint. arXiv:2012.01149
16. Jiang X, Wang S, Guo L, Wen Z, Jia L, Xu L, Xiao G, Li Q (2023) Integrating image and molecular profiles for spatial transcriptomics analysis. bioRxiv preprint. https://doi.org/10.1101/2023.06.18.545488
17. Wang S, Wang T, Yang L, YI F, Luo X, Yang Y, Gazdar A, Fujimoto J, Wistuba II, Yao B, Lin S, Xie Y, Mao Y, Xia, G (2018) ConvPath: A software tool for lung adenocarcinoma digital pathological image analysis aided by convolutional neural network. arXiv Preprint. arXiv:1809.10240
18. Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns, vol 70. Wiley, Chichester
19. Li Q, Yi F, Wang T, Xiao G, Liang F (2017) Lung cancer pathological image analysis using a hidden Potts model. Cancer Inf 16:1176935117711910
20. Li Q, Wang X, Liang F, Yi F, Xie Y, Gazdar A, Xiao G (2018) A Bayesian hidden Potts mixture model for analyzing lung cancer pathology images. Biostatistics 20(4):565–581

21. Hammersley JM, Clifford P (1971) Markov fields on finite graphs and lattices. Unpublished manuscript, Oxford University
22. Liang F (2010) A double Metropolis–Hastings sampler for spatial models with intractable normalizing constants. J Stat Comput Simul 80(9):1007–1022
23. Cox DR (1992) Regression models and life-tables. In: Breakthroughs in statistics. Springer, New York, pp 527–541
24. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, Vijver MJ, West RB, Rijn M, Koller D (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med 3(108):108–113
25. Wang S, Chen A, Yang L, Cai L, Xie Y, Fujimoto J, Gazdar A, Xiao G (2018) Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. Sci Rep 8(1):1–9
26. Fernández E, Yang S, Chiou SH, Moon C, Zhang C, Yao B, Xiao G, Li Q (2022) SAFARI: shape analysis for AI-segmented images. BMC Med Imaging 22(1):1–7
27. Kilday J, Palmieri F, Fox MD (1993) Classifying mammographic lesions using computerized image analysis. IEEE Trans Med Imaging 12(4):664–669
28. Georgiou H, Mavroforakis M, Dimitropoulos N, Cavouras D, Theodoridis S (2007) Multi-scaled morphological features for the characterization of mammographic masses using statistical classification schemes. Artif Intell Med 41(1):39–55
29. Li Z, Wang W, Shin S, Choi HD (2013) Enhanced roughness index for breast cancer benign/malignant measurement using Gaussian mixture model. In: Proceedings of the 2013 research in adaptive and convergent systems, pp 177–181
30. Rahmani Seryasat O, Haddadnia J, Ghayoumi Zadeh H (2016) Assessment of a novel computer aided mass diagnosis system in mammograms. Iran Q J Breast Dis 9(3):31–41
31. Sanghani P, Ti AB, King NKK, Ren H (2019) Evaluation of tumor shape features for overall survival prognosis in glioblastoma multiforme patients. Surg Oncol 29:178–183
32. Vadmal V, Junno G, Badve C, Huang W, Waite KA, Barnholtz-Sloan JS (2020) MRI image analysis methods and applications: an algorithmic perspective using brain tumors as an exemplar. Neuro-Oncol Adv 2(1):049
33. Strait J, Chkrebtii O, Kurtek S (2019) Automatic detection and uncertainty quantification of landmarks on elastic curves. J Am Stat Assoc 114(527):1002–1017
34. Meacham CE, Morrison SJ (2013) Tumour heterogeneity and cancer cell plasticity. Nature 501(7467):328–337
35. Dagogo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15(2):81
36. Moon C, Li Q, Xiao G (2020) Using persistent homology topological features to characterize medical images: case studies on lung and brain cancers. arxiv preprint. arXiv:2012.12102
37. Wang S, Rong R, Yang DM, Fujimoto J, Yan S, Cai L, Yang L, Luo D, Behrens C, Parra ER et al (2020) Computational staining of pathology images to study the tumor microenvironment in lung cancer. Cancer Res 80(10):2056–2066
38. Thrane K, Eriksson H, Maaskola J, Hansson J, Lundeberg J (2018) Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. Cancer Res 78(20):5970–5979
39. Moses L, Pachter L (2022) Museum of spatial transcriptomics. Nat Methods 19(5):534–546
40. Satija R, Farrell JA, Gennert D, Schier AF, Regev A (2015) Spatial reconstruction of single-cell gene expression data. Nat Biotechnol 33(5):495–502
41. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20(5):273–282
42. Zhu Q, Shah S, Dries R, Cai L, Yuan G-C (2018) Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. Nat Biotechnol 36(12):1183–1190
43. Pham D, Tan X, Xu J, Grice LF, Lam PY, Raghubar A, Vukovic J, Ruitenberg MJ, Nguyen Q (2020) stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell–cell interactions and spatial trajectories within undissociated tissues. bioRxiv preprint. https://doi.org/10.1101/2020.05.31.125658
44. Zhao E, Stone MR, Ren X, Guenthoer J, Smythe KS, Pulliam T, Williams SR, Uytingco CR, Taylor SE, Nghiem P et al (2021) Spatial transcriptomics at subspot resolution with BayesSpace. Nat Biotechnol 39(11):1375–1384

45. Hu J, Li X, Coleman K, Schroeder A, Ma N, Irwin DJ, Lee EB, Shinohara RT, Li M (2021) SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nat Methods 18(11):1342–1351
46. Fox H (2000) Is H &E morphology coming to an end? J Clin Pathol 53(1):38–40