Bayesian Libby-Novick and McDonald's Beta Mixture Models with Variational Inference

Diaa Azzam

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

November 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Diaa Azzam**

Entitled: **Bayesian Libby-Novick and McDonald's Beta Mixture Models with Variational Inference**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Abdelhak Bentaleb

_____ Examiner
Dr. Abdelhak Bentaleb

_____ Examiner
Dr. Zachary Patterson

_____ Supervisor
Dr. Nizar Bouguila

Approved by _____
Paquet, Joey, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Bayesian Libby-Novick and McDonald's Beta Mixture Models with Variational Inference

Diaa Azzam

Clustering is a foundational paradigm in data mining and pattern recognition, which is aimed at grouping and uncovering meaningful clusters. Clustering techniques are essential for extracting meaningful structure from data across a wide range of scientific domains. One of the main challenges is the clustering of bounded data that may not follow a Gaussian distribution and has an unknown number of clusters. Furthermore, the presence of irrelevant features poses fundamental challenges under the unsupervised learning setting. The aforementioned challenges undermine both cluster quality and obscure the downstream decision-making. This thesis aims to address these challenges by proposing three frameworks that leverage generalizations of the standard Beta distribution. First, we propose a Bayesian Libby-Novick Beta mixture model (BLNBMM) with integrated feature selection. Second, we propose a Variational Infinite Libby-Novick Beta Mixture Model (VILNBMM). Third, we introduce a Neural Variational Inference for Infinite McDonald's Beta Mixture Model (NVI-IMBMM). To enable posterior inference in our proposed models, we develop multiple variational inference (VI) frameworks. These variational inference algorithms recast posterior estimation into the form of an optimization problem. The proposed algorithms were evaluated on different medical imaging datasets. To benchmark our models, we compared them against established probabilistic mixture models. Our experiments showcase that the proposed models can indeed capture complex class distributions in bounded data domains.

Acknowledgments

I would like to convey my heartfelt thanks and profound appreciation to my supervisor, Professor Nizar Bouguila. His unwavering encouragement, guidance, and steadfast support that have been instrumental throughout my Master's studies.

I also would like to extend my warmest thanks and deep appreciation to Dr. Muhammad Azam for his invaluable guidance, persistent encouragement, and tireless support. I am truly thankful for this invaluable experience.

Finally, to my family, thank you for your support throughout this journey. Your endless support has been my foundation. I'm also grateful to my friends Manar and Ammar for their encouragement. This thesis is devoted to acknowledging everyone whose direct or indirect contributions made this master's thesis possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Clustering forms a cornerstone of unsupervised learning, providing essential methodologies for discovering group structures within data across numerous scientific and engineering domains. The era of big data has increased the challenges associated with modeling the different types of data distributions [1, 2, 3]. These challenges manifest in different fields such as bioinformatics, medical imaging, and computer vision. This is specifically where the underlying data structures exhibit patterns that refrain from having a direct categorization [4, 5, 6]. One of the challenges is the difficulty of navigating multidimensional feature spaces. These spaces are populated with both discriminative and irrelevant features. Furthermore, the presence of irrelevant features can degrade the pattern recognition process itself [7, 8, 9]. Consequently, the development of modeling approaches that possess the capability of accurately representing and extracting meaningful insights has become central to research [10].

Mixture models have emerged as probabilistic frameworks for density estimation and clustering. This is done through representing data as weighted combinations of simpler subpopulations [11, 12]. This decomposition models data into interpretable subpopulations [13, 14] which provides further insights about the underlying data generation processes [15]. The expressive power that mixture models attain has led to widespread adoption in medical imaging, signal processing, and pattern recognition [16, 17]. Traditional Gaussian mixture models offer computational efficiency and also

are mathematically tractable [18, 19]. However, they face limitations when confronted with real-world data exhibiting asymmetry and boundedness [20, 21, 22].

The Beta distribution and its generalizations provide alternatives for bounded data modeling [23, 24, 25]. While standard Beta distributions provide flexibility for skewed patterns through shape parameters [26, 27], they struggle with extreme skewness [28]. The Libby-Novick Beta (LNB) distribution addresses the flexibility limitation by having an additional shape parameter. This parameter enhances control over skewness and kurtosis [29]. Moreover, the LNB distribution has demonstrated superior performance in handling bounded data within mixture model frameworks [30, 31, 32, 33]. Additionally, there exists another distribution that generalizes the traditional Beta distribution. The McDonald's Beta distribution generalizes the traditional Beta distribution with two more parameters. This yields a more flexible control over location, scale, skewness, and tail behavior, making it well-suited for medical imaging and proportional datasets with non-Gaussian characteristics [34].

Traditional mixture models require a pre-specification of the number of components a priori; this, in fact, is a limitation that researchers have tried to address [35, 36]. In this context, the Dirichlet process (DP) provides a nonparametric solution to extending finite mixtures to infinite ones. This is done through a variety of mechanisms such as the stick breaking process [37, 38, 39]. However, incorporating flexible Beta generalizations within this framework introduces computational complexities [40]. The isolation and selection of relevant features forms another challenge, where the irrelevant features may degrade the overall clustering quality [41]. Integrating feature selection with mixture modeling addresses both efficiency and accuracy by focusing resources on relevant attributes [42, 43, 44]. In the context of mixture models, the integration of unsupervised feature selection methods with the generalizations of the Beta distribution is yet to be fully explored.

Computational challenges in parameter estimation have driven substantial methodological innovation. The Expectation-Maximization (EM) algorithm has seen widespread use for mixture model parameter estimation [45, 46]. On the other hand, Markov Chain Monte Carlo (MCMC) methods offer alternatives but become prohibitive for large-scale, high-dimensional problems [47, 48, 49]. Variational inference (VI) methods have provided alternatives to the EM algorithm. This is due to the fact that generally, VI methods aim to reformulate inference as optimization [50, 51, 42, 52].

Variational inference provides methods to quantify parameter uncertainty through approximate posterior distributions. Also, enables tractable inference in hierarchical models [53, 54, 55]. These advantages are particularly valuable for complex mixture models requiring both parameter estimation and uncertainty quantification [56]. Despite advances in the use of flexible distributional models, nonparametric frameworks, and efficient inference techniques; Significant challenges remain in modeling bounded data. There remains a need for frameworks that not only address these challenges but also offer interpretable clustering models to support evidence-based decision-making.

1.2 Contributions

The primary objectives of this thesis are threefold: First, we develop two infinite mixture models based on the Libby-Novick Beta distribution. Second, we develop infinite mixture model with the McDonald beta distribution with feature selection. Third, we propose different variational inference algorithms for the proposed models. Our contributions are summarized as follows:

- **Variational Inference for the Bayesian Libby Novick Beta Mixture Model with Feature Selection**

  In this work, we propose a Bayesian Libby-Novick Beta mixture model (BLNBMM) with integrated feature selection. To enable posterior inference in our proposed hierarchical model, we develop a variational inference (VI) framework. We evaluate the proposed model on both synthetic data as well as medical imaging data. This work has been submitted and currently under review [57]

- **Nonparametric Variational Infinite Libby-Novick Beta Mixture Model for Medical Data Clustering**

  In this work, we propose a Variational Infinite Libby-Novick Beta Mixture Model (VILNBMM), a framework that automatically determines the optimal number of clusters in imaging data by combining Dirichlet process priors with the Libby-Novick Beta (LNB) distribution. Our approach handles complex distributional patterns in bounded domains without requiring the presetting of component numbers a priori. We develop a variational inference algorithm that

reformulates posterior estimation as an optimization problem, making the model computationally tractable for real world applications. We evaluate this model on multiple medical imaging datasets. This work has been published in [58].

- **Nonparametric Neural Variational Inference for McDonald's Beta Mixture Models with Feature Selection**

  In this work, we introduce a Neural Variational Inference for Infinite McDonald's Beta Mixture Model (NVI-IMBMM). A framework that integrates neural variational inference with the infinite multivariate McDonald's Beta mixture model for simultaneous clustering and feature selection. Our approach leverages the flexibility of McDonald's Beta distributions to capture asymmetric patterns in bounded data while employing Dirichlet process to automatically determine the optimal number of clusters. We also assess our approach on medical imaging datasets. This work has been submitted and currently under review [59].

1.3 Thesis Overview

This thesis is organized as follows:

Chapter 1: In this chapter, we provide a background to the work presented in this thesis. We also define the contributions.

Chapter 2: We propose Bayesian Libby-Novick Beta mixture model with integrated feature selection. To enable posterior inference in our proposed hierarchical model, we develop a variational inference (VI) framework. We include experimental evaluation over medical images datasets.

Chapter 3: In this chapter, we introduce a Variational Infinite Libby-Novick Beta Mixture Model, a framework that automatically determines the optimal number of clusters in medical imaging data by combining Dirichlet process priors with the Libby-Novick Beta (LNB) distribution.

Chapter 4: The presented study in this chapter proposes a Neural Variational Inference for

Infinite McDonald's Beta Mixture Model. Here, we propose a framework that integrates neural variational inference with the infinite multivariate McDonald's Beta mixture model that performs both clustering and feature selection.

Chapter 5: In this chapter, we conclude our contributions and include closing remarks.

Chapter 2

Variational Inference for the Bayesian Libby-Novick Beta Mixture Model with Feature Selection

In this chapter, we present the full specification of our Bayesian Libby-Novick Beta mixture model with feature selection. We begin by defining the Libby-Novick Beta (LNB) distribution for each feature and show how it generalizes the standard Beta through an additional flexibility parameter. We then describe how observations are generated from a finite mixture of these LNB components, introduce latent indicator variables for both cluster membership and feature relevance, and detail the likelihood function that jointly accounts for relevant and irrelevant features. Finally, we embed this mixture model within a hierarchical Bayesian framework by specifying priors over all model parameters and hyperparameters, thereby enabling uncertainty quantification and regularization. Subsequent sections develop a variational inference algorithm to enable posterior approximation under this complex structure. The probability density function (PDF) of the LNB distribution for a single observation and according to the literature, is defined as shown in Equation 1. The LNB distribution is visualized in Fig.2.1.

$$p(x_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl}) = \frac{\lambda_{jl}^{\alpha_{jl}} \, x_{il}^{\alpha_{jl}-1} \, (1 - x_{il})^{\beta_{jl}-1}}{B(\alpha_{jl}, \beta_{jl}) \, [1 - (1 - \lambda_{jl}) x_{il}]^{\alpha_{jl}+\beta_{jl}}} \tag{1}$$

where $B(\alpha_{jl}, \beta_{jl})$ is the Beta function, serving as a normalization constant; $\alpha_{jl} > 0$ and $\beta_{jl} > 0$ are shape parameters controlling the distribution's form; $\lambda_{jl} \in (0, 1)$ provides additional flexibility in modeling skewness and kurtosis; and $x_{il} \in (0, 1)$ represents the observed value for data point i and feature l. It is also worth noting that, if λ_{jl} is set to 1, the LNB distribution reduces to the standard Beta distribution. The LNB distribution generalizes the Beta distribution by introducing an additional shape parameter λ, which enhances its capacity to model skewed and heavy-tailed data more flexibly than the standard Beta.

$$B(\alpha_{jl}, \beta_{jl}) = \int_0^1 t^{\alpha_{jl}-1}(1-t)^{\beta_{jl}-1}\,dt = \frac{\Gamma(\alpha_{jl})\Gamma(\beta_{jl})}{\Gamma(\alpha_{jl}+\beta_{jl})}. \tag{2}$$

Given a dataset $X = \{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$, where each $\mathbf{X}_i = (X_{i1}, \ldots, X_{iD})$ is a D-dimensional observation, we model each $\mathbf{X}_i$ as being generated from a finite mixture of M LNB distributions:

$$p(\mathbf{X}_i \mid \boldsymbol{\pi}, \boldsymbol{\theta}) = \sum_{j=1}^{M} \pi_j\, p(\mathbf{X}_i \mid \boldsymbol{\theta}_j) \tag{3}$$

Here, π_j represents the mixing coefficient for component j, satisfying $\sum_{j=1}^{M} \pi_j = 1$ and $\pi_j \geq 0$. The parameter set $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \boldsymbol{\lambda}_j)$ encapsulates the parameters for component j, and $p(\mathbf{X}_i \mid \boldsymbol{\theta}_j) = \prod_{l=1}^{D} p(X_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl})$, assuming conditional independence across features. To facilitate cluster assignments, we introduce a latent binary variable $z_{ij} \in \{0, 1\}$, where $z_{ij} = 1$ if data point i belongs to cluster j and $z_{ij} = 0$ otherwise, with the constraint $\sum_{j=1}^{M} z_{ij} = 1$ for each i. For feature selection, we incorporate a binary latent variable $\phi_{il} \in \{0, 1\}$, where $\phi_{il} = 1$ indicates that feature l is relevant for data point i, and $\phi_{il} = 0$ indicates that feature l is irrelevant. The probability of feature relevance is defined as $p(\phi_{il} = 1) = \rho_l$, where $\rho_l \in (0, 1)$ may vary across features. For irrelevant features ($\phi_{il} = 0$), we model them using a mixture of K Beta distributions with component assignments $w_{ilk} \in \{0, 1\}$, satisfying $\sum_{k=1}^{K} w_{ilk} = 1$ when $\phi_{il} = 0$. For a given dataset, the simplified likelihood is formulated as per Equation 4.

$$p(X \mid Z, \phi, W, \alpha, \beta, \lambda, \psi, \xi)$$
$$= \prod_{i=1}^{N} \prod_{l=1}^{D} \left(\sum_{j=1}^{M} z_{ij} \cdot \mathrm{LNB}\left(X_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl}\right) \right)^{\phi_{il}} \times \left(\sum_{k=1}^{K} w_{ilk} \cdot \mathrm{Beta}\left(X_{il} \mid \psi_{kl}, \xi_{kl}\right) \right)^{1-\phi_{il}} \tag{4}$$

where $\mathrm{LNB}(X_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl})$ is the LNB PDF for relevant features, $\mathrm{Beta}(X_{il} \mid \psi_{kl}, \xi_{kl})$ is the Beta PDF for irrelevant features, and $\psi_{kl} > 0$ and $\xi_{kl} > 0$ are shape parameters for the Beta distribution. This formulation ensures that relevant features contribute to clustering via the LNB distribution, while irrelevant features are modeled separately. For the irrelevant-feature Beta mixture, the parameters are denoted by $\boldsymbol{\psi} = \{\psi_{kl}\}$ and $\boldsymbol{\xi} = \{\xi_{kl}\}$ for $k = 1, \ldots, K$ and $l = 1, \ldots, D$. This model operates under the constraints that $X_{il} \in (0, 1)$ for all i and l. Additionally, the condition $\sum_{j=1}^{M} z_{ij} = 1$ must hold for all i. For i and l where $\phi_{il} = 0$, the constraint $\sum_{k=1}^{K} w_{ilk} = 1$ applies. We use l to index features throughout the model, regardless of feature relevance. Finally, the parameters satisfy $\sum_{k=1}^{K} v_k = 1$ with $v_k \geq 0$ for all k. Our model treats relevant and irrelevant features differently, and this design is motivated by intuitive information-theoretic reasoning. We assume that relevant features carry meaningful patterns that help distinguish between clusters. These patterns are often complex and require flexible distributions, such as the Libby-Novick Beta (LNB), to model them. In contrast, irrelevant features mostly reflect noise or information unrelated to the clustering structure. These can be represented well using simpler Beta distributions. This distinction reflects the idea that clustering increases similarity within groups and highlights differences between them, making the richer LNB model suitable for relevant features. In addition, it is worth clarifying that, in contrast to z_{ij}, which takes values in $\{0, 1\}$, the posterior $\hat{z}_{ij}$ which represents the soft assignment probability, is a continuous value in the interval $[0, 1]$

$$\hat{z}_{ij} = p(z_{ij} = 1 \mid \mathbf{X}_i, \boldsymbol{\theta}) = \frac{\pi_j \, p(\mathbf{X}_i \mid \boldsymbol{\theta}_j)}{\sum_{j'=1}^{M} \pi_{j'} \, p(\mathbf{X}_i \mid \boldsymbol{\theta}_{j'})} \tag{5}$$

This probability-based assignment allows each data point to belong to multiple clusters with varying probabilities rather than being deterministically assigned to a single cluster.

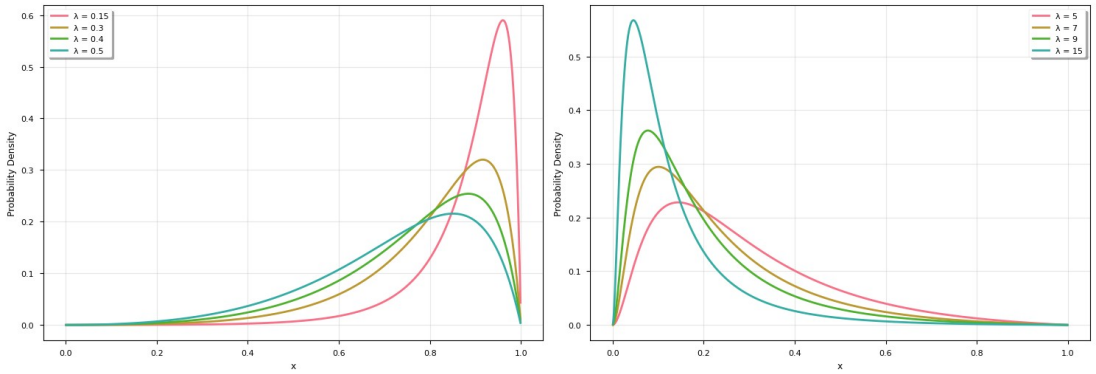

Figure 2.1: Visualization of the Libby-Novick Beta distribution exhibiting various shapes achieved through different λ values. The additional shape parameter λ provides enhanced flexibility in modeling asymmetric and heavy-tailed distributions compared to the standard Beta distribution.

2.1 Bayesian Hierarchical Structure

This section establishes a comprehensive Bayesian hierarchical framework for the BLNBMM (Bayesian Libby-Novick Beta Mixture Model) to characterize parameter uncertainty. This approach simultaneously quantifies uncertainty, enables uncertainty estimation, and introduces natural regularization. We define prior distributions that respect model constraints and enables posterior approximation for our model through variational methods, which we establish in section 2.2. For the Libby-Novick Beta distribution parameters, we employ Gamma priors that enforce the required positivity constraints:

$$\begin{aligned} p(\alpha_{jl}) &= \text{Gamma}(\alpha_{jl} \mid a_\alpha, b_\alpha), \\ p(\beta_{jl}) &= \text{Gamma}(\beta_{jl} \mid a_\beta, b_\beta), \\ p(\lambda_{jl}) &= \text{Beta}(\lambda_{jl} \mid a_\lambda, b_\lambda) \end{aligned} \tag{6}$$

These hyperparameters (a_α, b_α), (a_β, b_β), and (a_λ, b_λ) can be tuned to encode prior knowledge about the shape and scale of these distributions, or set to yield weakly informative priors when such knowledge is unavailable. For the background component parameters that model irrelevant features through Beta distributions, we adopt the priors as per Equation 7, additionally for the mixing proportions, which must satisfy the constraints, we utilize Dirichlet distributions as per Equation 8:

$$p(\psi_{kl}) = \text{Gamma}(\psi_{kl} \mid c_\psi, d_\psi), \quad p(\xi_{kl}) = \text{Gamma}(\xi_{kl} \mid c_\xi, d_\xi) \tag{7}$$

$$p(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\pi} \mid \boldsymbol{\gamma}), \qquad p(\boldsymbol{\nu}) = \text{Dirichlet}(\boldsymbol{\nu} \mid \boldsymbol{\delta}) \tag{8}$$

The concentration parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_M)$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$ control the prior beliefs about the relative prevalence of each mixture component. Values close to one yield nearly uniform distributions, in contrast to larger values which concentrate probability mass toward the expected proportions. To model feature relevance probabilities, which govern the inclusion/exclusion of features in the clustering mechanism, we employ Beta priors:

$$p(\rho_l) = \text{Beta}(\rho_l \mid a_\rho, b_\rho) \tag{9}$$

The hyperparameters a_ρ and b_ρ encode prior beliefs about feature relevance. Setting $a_\rho < b_\rho$ creates a sparsity-inducing prior that favors exclusion, aligning with the assumption that many features in the feature space maybe irrelevant to clustering. Each prior distribution has been chosen to facilitate reparameterizable variational approximations using the same distributional family. For example, Gamma and Beta priors are paired with Gamma and Beta variational posteriors to maintain conjugacy where possible, and ensure valid support for gradient-based optimization via the reparameterization trick. The complete joint distribution is factorized hierarchically according to the dependency structure of the model as:

$$p(X, Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\rho}) = \underbrace{p(X \mid Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi})}_{\text{Likelihood}}$$

$$\times \underbrace{p(Z \mid \boldsymbol{\pi}) \cdot p(\boldsymbol{\phi} \mid \boldsymbol{\rho}) \cdot p(W \mid \boldsymbol{\nu}, \boldsymbol{\phi})}_{\text{Latent variable distributions}} \quad \times \underbrace{p(\boldsymbol{\pi}) \cdot p(\boldsymbol{\nu}) \cdot p(\boldsymbol{\rho})}_{\text{Mixing proportion priors}} \cdot \underbrace{p(\boldsymbol{\alpha}) \cdot p(\boldsymbol{\beta}) \cdot p(\boldsymbol{\lambda}) \cdot p(\boldsymbol{\psi}) \cdot p(\boldsymbol{\xi})}_{\text{Distribution parameter priors}}$$

Expanding this joint distribution reveals the full probabilistic dependencies:

$$
\begin{aligned}
& p(X, Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\rho}) \\
&= \prod_{i=1}^{N} \prod_{l=1}^{D} \left[\prod_{j=1}^{M} \mathrm{LNB}(x_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl})^{z_{ij}} \right]^{\phi_{il}} \times \left[\sum_{k=1}^{K} (\nu_k \, \mathrm{Beta}(x_{il} \mid \psi_{kl}, \xi_{kl}))^{w_{ilk}} \right]^{1-\phi_{il}} \\
&\times \prod_{i=1}^{N} \prod_{j=1}^{M} \pi_j^{z_{ij}} \prod_{i=1}^{N} \prod_{l=1}^{D} \rho_l^{\phi_{il}} (1-\rho_l)^{1-\phi_{il}} \times \prod_{i=1}^{N} \prod_{l=1}^{D} \left[\sum_{k=1}^{K} \nu_k^{w_{ilk}} \right]^{1-\phi_{il}} \\
&\times \prod_{j=1}^{M} \prod_{l=1}^{D} p(\alpha_{jl} \mid a_\alpha, b_\alpha) \, p(\beta_{jl} \mid a_\beta, b_\beta) \, p(\lambda_{jl} \mid a_\lambda, b_\lambda) \times \prod_{k=1}^{K} \prod_{l=1}^{D} p(\psi_{kl} \mid c_\psi, d_\psi) \, p(\xi_{kl} \mid c_\xi, d_\xi) \\
&\times p(\boldsymbol{\pi} \mid \boldsymbol{\gamma}) \, p(\boldsymbol{\nu} \mid \boldsymbol{\delta}) \prod_{l=1}^{D} p(\rho_l \mid a_\rho, b_\rho)
\end{aligned}
\tag{10}
$$

This hierarchical structure offers several advantages. First, it automatically regularizes the model through the prior distributions, preventing degenerate solutions common in mixture modeling. Second, it enables feature selection by explicitly modeling relevance probabilities. Third, it provides a coherent framework for quantifying uncertainty in both cluster assignments and parameter estimates. Finally, the conditional independence assumptions embedded in this factorization, while simplifying computation, still capture the important dependencies between data points, features, and cluster assignments necessary for effective clustering. An illustrative figure of the BLNBMM can be found in Figure 2.2.

2.2 Variational Inference

Exact posterior inference for the BLNBMM is computationally intractable due to the complex dependencies between latent variables and the non-conjugate nature of the Libby-Novick Beta distribution parameters. The joint posterior distribution over all latent variables and parameters involves integrals that cannot be computed analytically. To address this challenge, we develop a variational inference framework that approximates the intractable posterior with a tractable surrogate

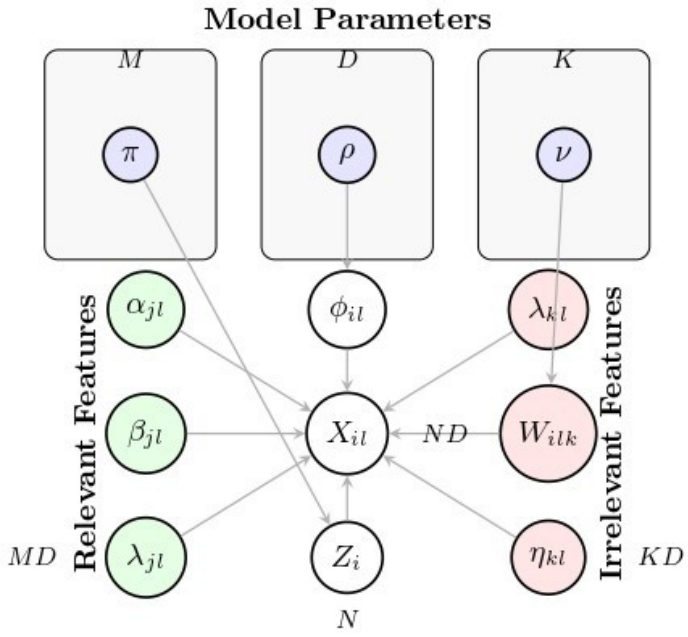

Figure 2.2: Graphical model representation of the Bayesian LNBMM with feature selection. Circles denote random variables. ϕ_{il} represents feature relevance indicators, Z_i cluster assignments, and W_{ilk} irrelevant feature component assignments. Boxes indicate repetition with labels showing dimensions: M : clusters, D : features, K : irrelevant components, N : data points. MD , ND , and KD indicate repetition over respective dimensions. Arrows show conditional dependencies

distribution, thus enabling inference. Our approach maintains the full Bayesian treatment while enabling computational tractability. We derive a coordinate ascent algorithm that alternates between updating the latent variable expectations (E-step) and optimizing the variational parameters (M-step). We approximate the intractable posterior distribution $p(Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\rho} \mid X)$ using the mean-field factorized variational distribution [60], which is shown in Equation 11. For conjugate parameter-prior pairs, we employ natural exponential family approximations that preserve conjugacy properties. The mixing proportions follow Dirichlet distributions as specified in Equation 12 and feature relevance probabilities are approximated by Beta distributions according to Equation 13. Additionally, for the LNB parameters $(\alpha_{jl}, \beta_{jl}, \lambda_{jl})$, we employ Gaussian approximations in transformed parameter spaces to ensure proper support constraints, as shown in Equation 14.

$$q(Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\pi}, \boldsymbol{\nu}, \boldsymbol{\rho}) = \prod_{i=1}^{N} q(\mathbf{z}_i) \prod_{i=1}^{N} \prod_{l=1}^{D} q(\phi_{il}) \times \prod_{i=1}^{N} \prod_{l=1}^{D} q(\mathbf{w}_{il}) \prod_{j=1}^{M} \prod_{l=1}^{D} q(\alpha_{jl}) q(\beta_{jl}) q(\lambda_{jl}) \times \prod_{k=1}^{K} \prod_{l=1}^{D} q(\psi_{kl}) q(\xi_{kl}) \times q(\boldsymbol{\pi}) q(\boldsymbol{\nu}) \prod_{l=1}^{D} q(\rho_l) \tag{11}$$

$$q(\boldsymbol{\pi}) = \text{Dirichlet}(\boldsymbol{\pi} \mid \tilde{\boldsymbol{\gamma}}), \quad q(\boldsymbol{\nu}) = \text{Dirichlet}(\boldsymbol{\nu} \mid \tilde{\boldsymbol{\delta}}) \tag{12}$$

$$q(\rho_l) = \text{Beta}(\rho_l \mid \tilde{a}_{\rho,l}, \tilde{b}_{\rho,l}) \tag{13}$$

$$\begin{aligned} q(\log \alpha_{jl}) &= N(\log \alpha_{jl} \mid \mu_{\alpha,jl}, \sigma^2_{\alpha,jl}) \\ q(\log \beta_{jl}) &= N(\log \beta_{jl} \mid \mu_{\beta,jl}, \sigma^2_{\beta,jl}) \\ q(\text{logit } \lambda_{jl}) &= N(\text{logit } \lambda_{jl} \mid \mu_{\lambda,jl}, \sigma^2_{\lambda,jl}) \end{aligned} \tag{14}$$

where $\text{logit}(x) = \log(x/(1-x))$ is the logit transformation. This parameterization automatically enforces the constraints $\alpha_{jl}, \beta_{jl} > 0$ and $\lambda_{jl} \in (0, 1)$ while enabling standard Gaussian variational inference techniques. Similarly, for the Beta distribution parameters of irrelevant features, we employ the log-normal approximations given in Equation 15.

The latent variable distributions are approximated using categorical and Bernoulli distributions as defined in Equation 16. Importantly, the feature relevance indicators ϕ_{il} are binary latent variables in the model, but we work with their posterior probabilities $\tilde{\phi}_{il} \in [0, 1]$ during variational inference, with final binary decisions obtained through thresholding.

$$\begin{aligned} q(\log \psi_{kl}) &= N(\log \psi_{kl} \mid \mu_{\psi,kl}, \sigma^2_{\psi,kl}) \\ q(\log \xi_{kl}) &= N(\log \xi_{kl} \mid \mu_{\xi,kl}, \sigma^2_{\xi,kl}) \end{aligned} \tag{15}$$

$$\begin{aligned} q(\mathbf{z}_i) &= \text{Categorical}(\mathbf{z}_i \mid \boldsymbol{\pi}_i) \\ q(\phi_{il}) &= \text{Bernoulli}(\phi_{il} \mid \tilde{\phi}_{il}) \\ q(\mathbf{w}_{il}) &= \text{Categorical}(\mathbf{w}_{il} \mid \boldsymbol{\nu}_{il}) \end{aligned} \tag{16}$$

where $\boldsymbol{\pi}_i = (\tilde{\pi}_{i1}, \ldots, \tilde{\pi}_{iM})$ represents the variational cluster assignment probabilities, $\tilde{\phi}_{il}$ denotes the variational feature relevance probability, and $\tilde{\boldsymbol{\nu}}_{il}$ represents the variational irrelevant component assignment probabilities.

2.2.1 Evidence Lower Bound

The evidence lower bound (ELBO) provides an objective function for variational optimization. By Jensen's inequality, we derive the fundamental bound shown in Equation 17, where $\boldsymbol{\Theta}$ collectively denotes all latent variables and parameters, and $L(q)$ is the ELBO. Expanding this bound yields the decomposition in Equation 18.

$$\begin{aligned} \log p(X) &= \log \int_{Z} p(X, \boldsymbol{\Theta}) d\boldsymbol{\Theta} \\ &= \log \int_{Z} \frac{p(X, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} q(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \\ &\geq \int_{Z} q(\boldsymbol{\Theta}) \log \frac{p(X, \boldsymbol{\Theta})}{q(\boldsymbol{\Theta})} d\boldsymbol{\Theta} = L(q) \end{aligned} \tag{17}$$

$$L(q) = E_q[\log p(X, \boldsymbol{\Theta})] - E_q[\log q(\boldsymbol{\Theta})] = E_q[\log p(X, \boldsymbol{\Theta})] + H[q(\boldsymbol{\Theta})] \tag{18}$$

The first term represents the expected complete data log-likelihood under the variational distribution, while the second term $H[q(\boldsymbol{\Theta})]$ is the entropy of the variational distribution. We can further decompose the ELBO into its constituent components as presented in Equation 19. The expected data log-likelihood term, which is central to our inference procedure, becomes as expressed in Equation 20 and Equation 21 is the expectation computation for the LNB distribution.

$$\begin{aligned} L(q) = E_q &\log p(X \mid Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}) \\ &+ E_q \log p(Z \mid \boldsymbol{\pi}) + E_q \log p(\boldsymbol{\phi} \mid \boldsymbol{\rho}) + E_q \log p(W \mid \boldsymbol{\nu}) \\ &+ E_q \log p(\boldsymbol{\pi}) + E_q \log p(\boldsymbol{\nu}) + E_q \log p(\boldsymbol{\rho}) \\ &+ E_q \log p(\boldsymbol{\alpha}) + E_q \log p(\boldsymbol{\beta}) + E_q \log p(\boldsymbol{\lambda}) \\ &+ E_q \log p(\boldsymbol{\psi}) + E_q \log p(\boldsymbol{\xi}) + H[q] \end{aligned} \tag{19}$$

$$E_q \log p(X \mid \Theta) = \sum_{i=1}^{N} \sum_{\ell=1}^{D} \Big(\tilde{\phi}_{i\ell} \sum_{j=1}^{M} \pi_{ij} E_q \log \mathrm{LNB}(X_{i\ell} \mid \alpha_{j\ell}, \beta_{j\ell}, \lambda_{j\ell}) + (1 - \tilde{\phi}_{i\ell}) \sum_{k=1}^{K} \nu_{i\ell k} E_q \log \mathrm{Beta}(X_{i\ell} \mid \psi_{k\ell}, \xi_{k\ell}) \Big) \quad (20)$$

where $\Theta = \{Z, \boldsymbol{\phi}, W, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\psi}, \boldsymbol{\xi}\}$.

$$E_q \log \mathrm{LNB}(X_{i\ell} \mid \alpha_{j\ell}, \beta_{j\ell}, \lambda_{j\ell}) \approx \bar{\alpha}_{j\ell} \log \bar{\lambda}_{j\ell} + (\bar{\alpha}_{j\ell} - 1) \log X_{i\ell} + (\bar{\beta}_{j\ell} - 1) \log(1 - X_{i\ell}) - E_q[\log B(\alpha_{j\ell}, \beta_{j\ell})] - (\bar{\alpha}_{j\ell} + \bar{\beta}_{j\ell}) \log(1 - (1 - \bar{\lambda}_{j\ell}) X_{i\ell}) \quad (21)$$

For the log-normal and logit-normal variational posteriors, we compute the required expectations. The computations for log-normal parameters are given by Equation 22. However, for the logit-normal parameter λ_{jl}, we employ the delta method approximation. Let $\lambda_{jl} = \mathrm{sigmoid}(\mu_{\lambda,jl} + \sigma_{\lambda,jl}\,\varepsilon)$ where $\varepsilon \sim N(0, 1)$. The approximation in Equation 23 is valid when $\sigma^2_{\lambda,jl}$ is small relative to $\mu^2_{\lambda,jl}$, which we enforce during optimization. For the terms involving products of random variables, we employ as per Equation 25.

$$\begin{aligned} E_q[\alpha_{jl}] &= \exp\left(\mu_{\alpha,jl} + \frac{1}{2}\sigma^2_{\alpha,jl}\right), \\ E_q[\beta_{jl}] &= \exp\left(\mu_{\beta,jl} + \frac{1}{2}\sigma^2_{\beta,jl}\right), \\ E_q[\log \alpha_{jl}] &= \mu_{\alpha,jl}, \quad E_q[\log \beta_{jl}] = \mu_{\beta,jl} \end{aligned} \quad (22)$$

$$\begin{aligned} E_q[\lambda_{jl}] &\approx \mathrm{sigmoid}(\mu_{\lambda,jl}), \\ E_q[\log \lambda_{jl}] &\approx \log(\mathrm{sigmoid}(\mu_{\lambda,jl})) - \frac{\sigma^2_{\lambda,jl}}{2}\mathrm{sigmoid}(\mu_{\lambda,jl})\mathrm{sigmoid}(-\mu_{\lambda,jl}) \end{aligned} \quad (23)$$

where $\bar{\alpha}_{jl} = E_q[\alpha_{jl}]$, $\bar{\beta}_{jl} = E_q[\beta_{jl}]$, $\bar{\lambda}_{jl} = E_q[\lambda_{jl}]$, $\log \lambda_{jl} = E_q[\log \lambda_{jl}]$ (24)

$$\begin{aligned} E_q[\log B(\alpha_{jl}, \beta_{jl})] &= E_q[\log \Gamma(\alpha_{jl})] + E_q[\log \Gamma(\beta_{jl})] - E_q[\log \Gamma(\alpha_{jl} + \beta_{jl})] \\ E_q[\log \Gamma(\alpha_{jl})] &\approx \log \Gamma(E_q[\alpha_{jl}]) + \frac{1}{2}\psi'(E_q[\alpha_{jl}])\,\mathrm{Var}_q[\alpha_{jl}] \end{aligned} \quad (25)$$

where $\psi'(\cdot)$ is the trigamma function and the approximation is valid for posteriors. The delta method approximations are employed when the variational variances satisfy the conditions in Equation. 26. This condition is enforced during optimization by constraining variances to $[\sigma^2_{\min}, \sigma^2_{\max}]$ where $\sigma^2_{\max}$ ensures validity.

$$\sigma^2_{\alpha,jl}, \sigma^2_{\beta,jl}, \sigma^2_{\lambda,jl} \ll (\mu_{\alpha,jl}, \mu_{\beta,jl}, \mu_{\lambda,jl})^2 \tag{26}$$

2.2.2 Coordinate Ascent Optimization

We optimize the ELBO using a coordinate ascent algorithm that alternates between updating the latent variable expectations (E-step) and the parameter variational distributions (M-step). As for the E-step updates, the optimal variational distributions for the latent variables are obtained by setting the derivative of the ELBO to zero. The update rule for the cluster assignments is :

$$\log \pi_{ij} = E_q[\log \pi_j] + \sum_{l=1}^{D} \tilde{\phi}_{il} E_q[\log \text{LNB}(X_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl})] + C \tag{27}$$

where C is a normalization constant ensuring $\sum_j \pi_{ij} = 1$. For feature relevance indicators, we derive the sigmoid-based update as shown in Equation 28. Moreover, the irrelevant component assignments we have the update in Equation 29.

$$\begin{aligned} \tilde{\phi}_{il} &= \text{sigmoid}(\Delta_{il}) \\ \text{where } \Delta_{il} &= E_q[\log \rho] - E_q[\log(1-\rho_l)] + \sum_{j=1}^{M} \pi_{ij} E_q[\log \text{LNB}(X_{il} \mid \alpha_{jl}, \beta_{jl}, \lambda_{jl})] \\ &\quad - \sum_{k=1}^{K} \nabla_{ilk} E_q[\log \text{Beta}(X_{il} \mid \psi_{kl}, \xi_{kl})] \end{aligned} \tag{28}$$

$$\log \nabla_{ilk} = E_q[\log \nu_k] + E_q[\log \text{Beta}(X_{il} \mid \psi_{kl}, \xi_{kl})] + C \tag{29}$$

Subsequently, in the M-step updates concerning the conjugate parameters, we obtain updates as per Equations 30 and 31. Conversely, regarding the LNB parameters, we employ gradient-based updates. The gradient for the α parameter is as Equation 32, subsequently the β parameter gradient

and parameter λ_{jl} we have the following Equations 34 and 35 respectively, where the λ_{jl} gradient is in the logit space.

$$\bar{\gamma}_j = \gamma_j + \sum_{i=1}^{N} \pi_{ij} \qquad \bar{a}_{\rho,l} = a_\rho + \sum_{i=1}^{N} \tilde{\phi}_{il}$$
$$\bar{\delta}_k = \delta_k + \sum_{i=1}^{N} \sum_{l=1}^{D} (1 - \tilde{\phi}_{il}) \nu_{ilk} \qquad (30) \qquad \bar{b}_{\rho,l} = b_\rho + \sum_{i=1}^{N} (1 - \tilde{\phi}_{il}) \qquad (31)$$

$$\frac{\partial L}{\partial \mu_{\alpha,jl}} = \sum_{i=1}^{N} \pi_{ij} \tilde{\phi}_{il} \left[\log \lambda_{jl} + \log X_{il} - E_q[\psi(\alpha_{jl})] + E_q[\psi(\alpha_{jl} + \beta_{jl})] \right.$$
$$\left. - \log(1 - (1 - \bar{\lambda}_{jl}) X_{il}) \right] \bar{\alpha}_{jl} + \frac{\partial}{\partial \mu_{\alpha,jl}} E_q[\log p(\alpha_{jl})] \qquad (32)$$

$$E_q[\psi(\alpha_{jl})] \approx \psi(E_q[\alpha_{jl}]) + \frac{1}{2} \psi'(E_q[\alpha_{jl}]) \operatorname{Var}_q[\alpha_{jl}] \qquad (33)$$

$$\frac{\partial L}{\partial \mu_{\beta,jl}} = \sum_{i=1}^{N} \pi_{ij} \tilde{\phi}_{il} \left[\log(1 - X_{il}) - E_q[\psi(\beta_{jl})] + E_q[\psi(\alpha_{jl} + \beta_{jl})] \right.$$
$$\left. - \log(1 - (1 - \bar{\lambda}_{jl}) X_{il}) \right] \bar{\beta}_{jl} + \frac{\partial}{\partial \mu_{\beta,jl}} E_q[\log p(\beta_{jl})] \qquad (34)$$

$$\frac{\partial L}{\partial \mu_{\lambda,jl}} = \sum_{i=1}^{N} \pi_{ij} \tilde{\phi}_{il} \left[\bar{\alpha}_{jl} \frac{\partial \log \lambda_{jl}}{\partial \mu_{\lambda,jl}} - (\bar{\alpha}_{jl} + \bar{\beta}_{jl}) \frac{X_{il} \bar{\lambda}_{jl} (1 - \bar{\lambda}_{jl})}{1 - (1 - \bar{\lambda}_{jl}) X_{il}} \right.$$
$$\left. + \frac{\partial}{\partial \mu_{\lambda,jl}} E_q[\log p(\lambda_{jl})] \right] \qquad (35)$$

Parameter updates employ adaptive gradient ascent with momentum:

$$\mu_{\alpha,jl}^{(t+1)} = \mu_{\alpha,jl}^{(t)} + \eta_t \left[\beta m_{\alpha,jl}^{(t-1)} + (1 - \beta) \operatorname{clip}\left(\frac{\partial L}{\partial \mu_{\alpha,jl}}, -\tau, \tau \right) \right]$$
$$\sigma_{\alpha,jl}^{2\,(t+1)} = \operatorname{clip}\left(\sigma_{\alpha,jl}^{2\,(t)} + \eta_t \frac{\partial L}{\partial \sigma_{\alpha,jl}^2}, \sigma_{\min}^2, \sigma_{\max}^2 \right) \qquad (36)$$

where gradient clipping with threshold τ and variance clipping provide better numerical stability. Algorithm 1 summarizes the complete variational inference procedure.

Algorithm 1 Variational Inference for BLNBMM

Input: Data X, components M, irrelevant components K, hyperparameters
Initialize: Variational parameters using method of moments
Set variance bounds: $\sigma^2_{\min} = 10^{-6}$, $\sigma^2_{\max}$
$t \leftarrow 0$, $L^{(0)} \leftarrow -\infty$
repeat
    E-step:
        Update cluster assignments π_{ij} using Eq. 27
        Update feature relevance $\tilde{\phi}_{il}$ using Eq. 28
        Update irrelevant assignments ν_{ilk} using Eq. 29
    M-step:
        Update mixing parameters $\boldsymbol{\gamma}$, $\boldsymbol{\delta}$ using Eq. 30
        Update relevance parameters $\tilde{a}_{\rho,l}$, $\tilde{b}_{\rho,l}$ using Eq. 31
        Update LNB parameters using Eq. 36 with gradient clipping
        Enforce variance bounds per Eq. 26
        Update irrelevant parameters using corrected gradients
    Compute ELBO $L^{(t+1)}$ using Eq. 19
until Convergence per $t > T_{\max}$
Return: Optimized variational parameters

2.3 Experimental Evaluation

Our experimental framework evaluates the proposed BLNBMM through two approaches. First is on synthetic data, second is through real medical imaging applications. We present our experimental pipeline, with the complete pipeline shown in Figure 2.3. We selected diverse medical imaging datasets representing various domains with different characteristics and complexity levels. These datasets present challenging clustering test cases due to their variability and domain specific features, making them ideal for validating our proposed model on real world medical applications. Our preprocessing framework has three stages (Figure 2.3). The feature transformation pipeline employs Scale-Invariant Feature Transform (SIFT) [61] to extract local patterns while maintaining invariance to scaling. These extracted features are subsequently transformed through the Bag of Visual Words (BoVW) methodology [62], generating frequency histogram representations that effectively capture image content. To maintain statistical adherence and the bounded domain requirements of the LNB distribution, we perform normalization, which maps all values to the $[0, 1]$ interval. This step preserves discriminative power while ensuring compatibility with BLNBMM's distributional assumptions.

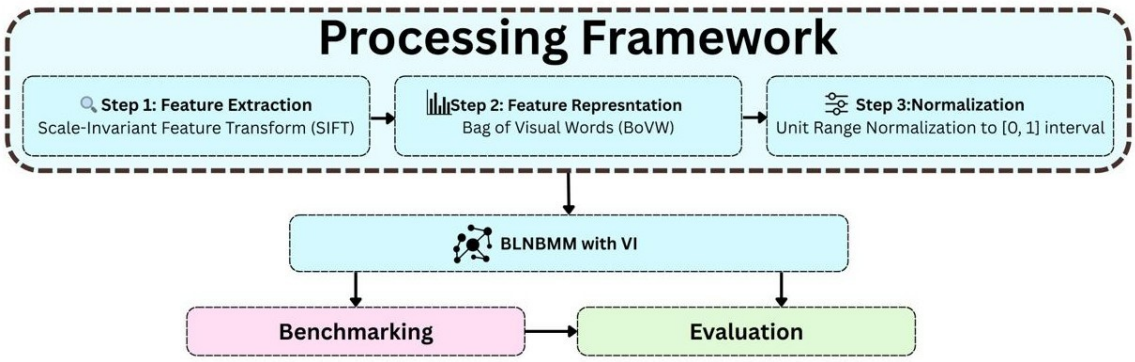

Figure 2.3: Experimental pipeline for the Bayesian Libby-Novick Beta Mixture Model (BLNBMM)

We benchmarked our BLNBMM with VI for a comprehensive analysis against an established probabilistic clustering approach. Namely, these are the classical Gaussian Mixture Model (GMM), the Variational Gaussian Mixture Model (VGMM), and the Beta Mixture Model (BMM). These baselines provide comparison across both frequentist and Bayesian paradigms, enabling us to quantify the specific advantages of the Libby-Novick Beta distribution in mixture modeling. We evaluated clustering performance on the real world data using four complementary metrics: Purity, Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and Fowlkes-Mallows Index (FMI). Purity quantifies cluster homogeneity as $\text{Purity} = \frac{1}{N}\sum_{i=1}^{k} \max_j |c_i \cap t_j|$, where N is the sample size, c_i is cluster i, and t_j is class j. The ARI measures clustering agreement adjusted for chance: $\text{ARI} = \frac{\text{RI} - E[\text{RI}]}{\max(\text{RI}) - E[\text{RI}]}$, where RI is the Rand Index. NMI evaluates shared information between clusters and true labels: $\text{NMI}(U, V) = \frac{\sqrt{I(U,V)}}{H(U)H(V)}$, with I as mutual information and H as entropy. Finally, FMI provides the geometric mean of precision and recall: $\text{FMI} = \sqrt{\text{PPV} \times \text{TPR}}$. This comprehensive evaluation suite allows us to assess clustering quality across multiple dimensions of the clustering landscape.

2.3.1 Artificial Data Evaluation

In this section, we evaluated our proposed BLNBMM-VI on synthetic datasets with known ground truth clustering structures and feature relevance patterns. Three datasets were generated with varying complexity to assess clustering accuracy, parameter recovery, and feature selection capabilities under controlled conditions. Each synthetic dataset contained features following LNB

distributions for relevant features and standard Beta distributions for irrelevant features. For relevant features, three distinct parameter regimes were implemented across clusters: high skewness toward 1 ($\alpha \in [5, 8]$, $\beta \in [1.5, 3]$, $\lambda \in [0.7, 0.9]$), high skewness toward 0 ($\alpha \in [1.5, 3]$, $\beta \in [5, 8]$, $\lambda \in [0.1, 0.3]$), and moderate distributions ($\alpha \in [3, 5], \beta \in [3, 5], \lambda \in [0.4, 0.6]$). Tables 2.1 and 2.2 presents the dataset specifications, parameter recovery results, and feature selection performance. Parameter recovery was assessed using error between true and estimated parameter means. Feature selection was evaluated based on the model's ability to truly identify relevant features.

Table 2.1: Evaluation Results on Synthetic Datasets

	Dataset Specifications				Identified
	Samples	Features	Clusters	True Rel. Features	Features
Test case 1	900	15	3	10	8
Test case 2	800	20	2	10	10
Test case 3	1500	30	3	12	9

Table 2.2: Evaluation Results on Synthetic Datasets – Mean Parameter Recovery

	α		β		λ	
	True	Est.	True	Est.	True	Est.
Test case 1	4.406	4.478	4.194	4.164	0.490	0.518
Test case 2	3.958	3.908	4.130	4.138	0.492	0.479
Test case 3	4.629	4.649	4.222	4.169	0.498	0.486

The results demonstrate the performance of our proposed model across synthetic datasets of varying complexity. Parameter recovery shows excellent accuracy, with mean absolute errors typically below 0.1 for all three parameters (α, β, λ). The model captured the distributional characteristics across clusters, with estimated parameters closely matching the true values even in challenging scenarios with overlapping parameter ranges. These synthetic data results showcase our method's ability to recover underlying cluster structures and identify relevant features when the true generative model aligns with our proposed BLNBMM framework.

2.3.2 Malaria Detection

Malaria poses a considerable challenge to global health and necessitates accurate diagnostic methods to enable effective treatment strategies. Traditionally, malaria diagnosis has relied on the

microscopic examination of stained blood smears [63]. In this process, trained clinicians identify infected cells based on their distinct morphological characteristics. This diagnostic technique is challenging, this was primarily due to the requirement for a thorough analysis of a high volume of cells to ensure accurate detection of the malaria parasite [64]. As illustrated in Figure 3.4, the diagnosis involves careful examination of blood smear images. In our study, we have addressed these challenges by employing a dataset comprising 1,400 images, which is balanced across the two classes, infected and uninfected cells. This dataset was obtained from the work of [65], providing a foundation for further research by utilizing this comprehensive collection of images.

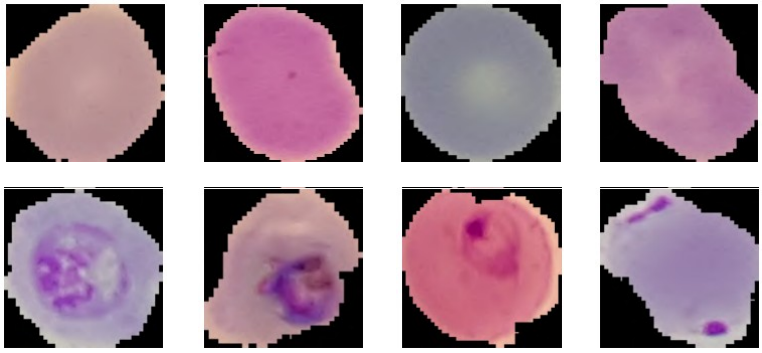

Figure 2.4: Examples of cell images. Upper row: uninfected (normal) cells. Bottom row: malaria-infected cells.

Table 2.3: Comparison of Clustering Methods for Malaria Dataset

Metric	Method			
	BLNBMM-VI	GMM	BMM	VGMM
Purity	0.904	0.876	0.887	0.879
ARI	0.652	0.461	0.485	0.474
NMI	0.544	0.474	0.491	0.483
FMI	0.826	0.743	0.752	0.748

Table 2.3 presents the clustering performance of various methods applied to the malaria dataset. Moreover Fig.2.5 showcases the most informative selected features, based on the feature selection mechanism. The proposed BLNBMM-VI approach demonstrates superior performance, achieving Purity (0.904), ARI (0.652), NMI (0.544), and FMI (0.826). These results indicate enhanced separation between infected and uninfected cells compared to traditional methods. Notably, BLNBMM-VI shows substantial improvements over the second best performing method BMM.

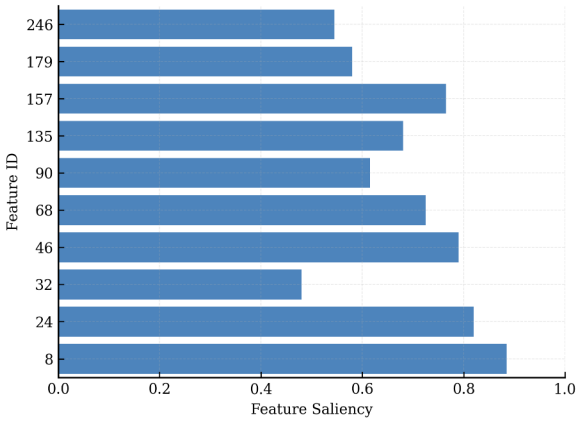

Figure 2.5: Overall Top 10 Most Relevant Features for Malaria Detection Dataset

2.3.3 Colon and Lung Cancer Diagnosis

Lung cancer remains a significant global health challenge, ranking as one of the leading causes of cancer related mortality [66]. Among its various subtypes, adenocarcinoma and squamous cell carcinoma are the most commonly diagnosed forms, requiring specialized diagnostic approaches for effective treatment planning [67]. Histopathological image analysis has emerged as a critical tool in accurately identifying and classifying these cancer subtypes, allowing for more targeted therapeutic interventions. Our study used the comprehensive dataset introduced by [68], which contains histopathological images across multiple tissue types. In this evaluation, we used the three lung cancer classes (adenocarcinoma, squamous cell carcinoma, and benign tissue). In addition, we also used two colon classes (adenocarcinoma and benign tissue). This diverse dataset provides an excellent foundation for evaluating our clustering methodology across different cancer types, enhancing the generalizability of our findings. We selected 500 samples from each class for both lung and colon categories to ensure balanced representation. Figure 2.7 illustrates representative samples from the lung dataset. Moreover, Figure 2.6 illustrates image samples drawn from the colon dataset.

Tables 2.4 and 2.5 present the clustering performance of various methods on colon and lung histopathology datasets, respectively. Across both datasets, the proposed BLNBMM-VI model outperformed conventional approaches. The results indicate strong class separation and accurate

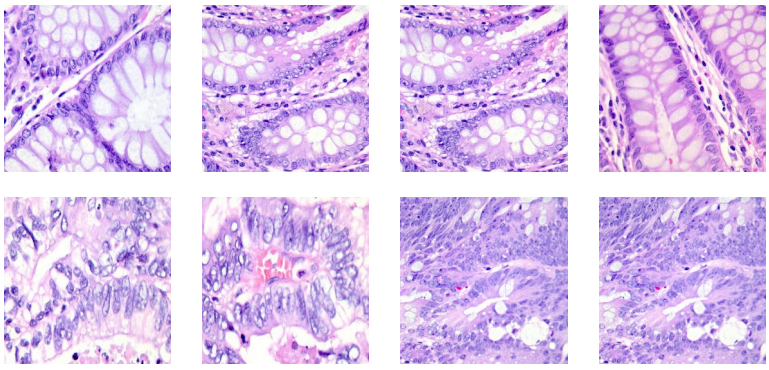

Figure 2.6: Examples of colon tissue images. Upper row: normal (benign) colon tissue samples. Bottom row: colon adenocarcinoma (cancerous) tissue samples.

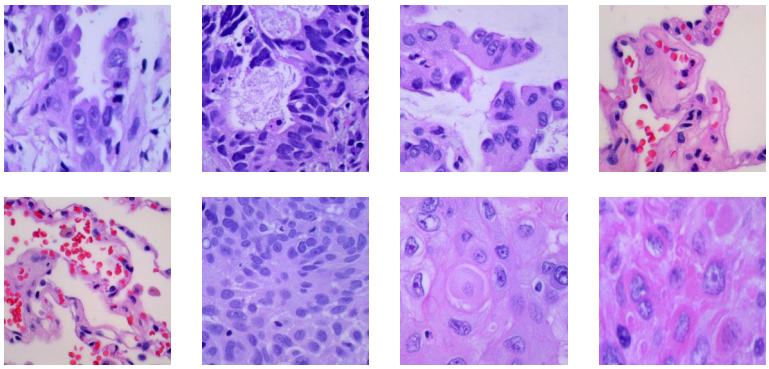

Figure 2.7: Examples of lung tissue images contianing a sequence of different tissue types such lung adenocarcinoma, followed by two images of normal (benign) lung tissue, and the final three images display lung squamous cell carcinoma.

Table 2.4: Comparison of Clustering Methods for Colon Dataset

Metric	Method			
	BLNBMM-VI	GMM	BMM	VGMM
Purity	0.929	0.885	0.891	0.882
ARI	0.719	0.591	0.630	0.586
NMI	0.681	0.485	0.517	0.492
FMI	0.866	0.795	0.809	0.790

Table 2.5: Comparison of Clustering Methods for Lung Dataset

Metric	Method			
	BLNBMM-VI	GMM	BMM	VGMM
Purity	0.862	0.849	0.829	0.842
ARI	0.650	0.618	0.580	0.600
NMI	0.661	0.571	0.558	0.556
FMI	0.759	0.746	0.718	0.735

grouping of cancerous versus benign tissues. A similar trend is observed in the lung dataset, where the model maintains a competitive edge with ARI. To further showcase the discriminative power of our approach, we analyzed the most relevant features identified by the BLNBMM-VI model for both datasets and this is visualized in Figures 2.8 and 4.3 respectively.

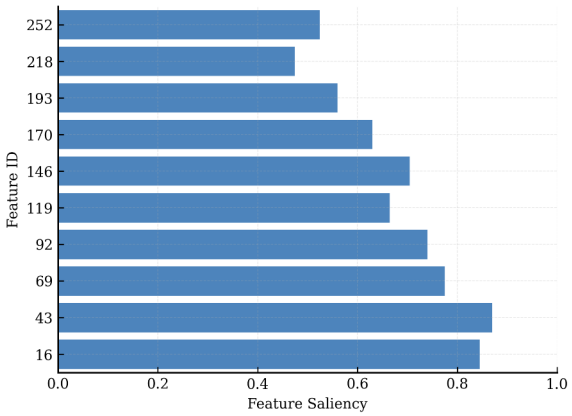

Figure 2.8: Overall Top 10 Relevant Features for Colon Cancer Dataset

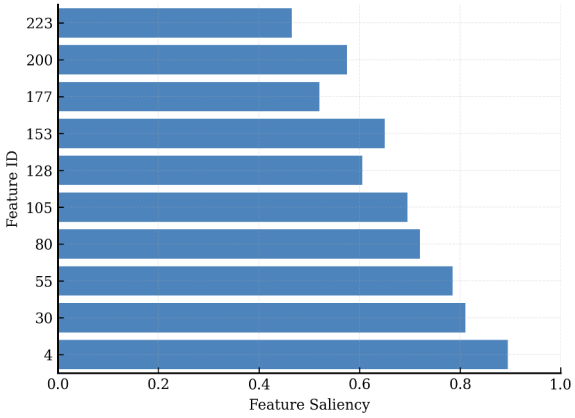

Figure 2.9: Overall Top 10 Relevant Features for Lung Cancer Dataset

Chapter 3

Nonparametric Variational Infinite Libby-Novick Beta Mixture Model for Medical Data Clustering

In this chapter, we provide a detailed exposition of our proposed Nonparametric Variational Infinite Libby Novick Beta Mixture Model. At the core of our approach is the Libby-Novick Beta (LNB) distribution, a generalization of the standard Beta distribution characterized by an additional shape parameter λ. For a random variable $x \in (0, 1)$ with parameters $a, b, \lambda > 0$, the probability density function is defined as:

$$p(x|a, b, \lambda) = \frac{\lambda^a x^{a-1} (1 - x)^{b-1}}{B(a, b)\{1 - (1 - \lambda)x\}^{a+b}} \tag{37}$$

where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ is the Beta function, and $\Gamma(\cdot)$ is the gamma function. The standard Beta distribution is recovered when $\lambda = 1$. The LNB distribution provides modeling flexibility through its three parameters, a and b control the basic shape similar to the standard Beta distribution, while the additional parameter λ simultaneously influences both skewness and kurtosis. This enhanced parameterization makes the LNB distribution particularly effective for modeling complex, asymmetric data patterns that appear in real-world applications. Fig.1 showcases the LNB distribution with varying control parameter values.

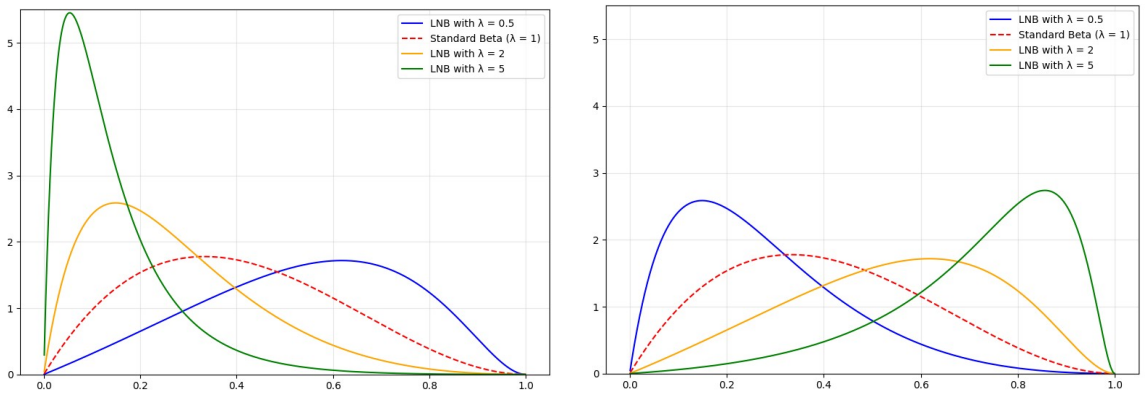

Figure 3.1: Comparison of LNB distribution. (left) with parameters $a = 2, b = 3$ for various λ values; (right) the corresponding inverse distribution with parameters $a = 3, b = 2$.

3.1 Mixture Model Formulation

Before introducing the infinite model, we define the finite Libby-Novick Beta mixture model. Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N independent and identically distributed D-dimensional vectors, where each component $x_{id} \in (0, 1)$ for $i = 1, \ldots, N$ and $d = 1, \ldots, D$. The finite LNB mixture model with K components is defined as:

$$p(X \mid \boldsymbol{\pi}, \boldsymbol{\Theta}) = \prod_{i=1}^{N} \left[\sum_{k=1}^{K} \pi_k p(\mathbf{x}_i \mid \boldsymbol{\theta}_k) \right] \tag{38}$$

$$p(\mathbf{x}_i \mid \boldsymbol{\theta}_k) = \prod_{d=1}^{D} \frac{\lambda_{kd}^{a_{kd}} x_{id}^{a_{kd}-1} (1 - x_{id})^{b_{kd}-1}}{B(a_{kd}, b_{kd}) \{1 - (1 - \lambda_{kd}) x_{id}\}^{a_{kd}+b_{kd}}} \tag{39}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ represents the mixing coefficients satisfying $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$. The parameter set $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K\}$ contains the distribution parameters for each component, with $\boldsymbol{\theta}_k = (\mathbf{a}_k, \mathbf{b}_k, \boldsymbol{\lambda}_k)$ where $\mathbf{a}_k = (a_{k1}, \ldots, a_{kD})$, $\mathbf{b}_k = (b_{k1}, \ldots, b_{kD})$, and $\boldsymbol{\lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kD})$, with $a_{kd}, b_{kd}, \lambda_{kd} > 0$ for all k and d.

To develop our infinite mixture model, we leverage the theory of Dirichlet processes [69]. We begin by stating the formulation of the Dirichlet distribution, a multivariate generalization of the Beta distribution that provides a distribution over a set of positive real numbers that sum to one. For a parameter vector $\alpha = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$, the probability density function of a Dirichlet

distribution over $\pi = (\pi_1, \ldots, \pi_K)$ with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$ is:

$$p(\pi|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{40}$$

The Dirichlet process (DP), denoted as DP (G_0, α), extends this concept to infinite dimensions, providing a distribution over distributions. It is characterized by a base distribution G_0 and a concentration parameter $\alpha > 0$. If $G \sim \text{DP}(G_0, \alpha)$, then for any finite partition $(A_1, \ldots, A_K)$ of the sample space:

$$(G(A_1), \ldots, G(A_K)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K)) \tag{41}$$

The concentration parameter α controls the variability around the base distribution G_0, with larger values producing distributions more similar to G_0. To compute this, we utilize the stick-breaking construction of the Dirichlet process. In this representation, a random probability measure G drawn from DP(G_0, α) can be expressed as $G = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$ where δ_{θ_j} is a point mass at θ_j, $\theta_j \overset{\text{i.i.d.}}{\sim} G_0$, and the weights π_j are constructed via $\pi_j = v_j \prod_{l=1}^{j-1} (1 - v_l)$ with $v_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha)$. This construction can be visualized as repeatedly breaking a unit-length stick. First breaking off a portion v_1 to get $\pi_1 = v_1$, then breaking a portion v_2 of the remainder $(1 - v_1)$ to get $\pi_2 = v_2(1 - v_1)$, and continuing indefinitely to yield an infinite sequence of weights summing to 1. Now we define the generative model, using this framework, we model a D-dimensional dataset $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iD}) \in (0, 1)^D$ as a Dirichlet process mixture of multivariate LNB distributions. The generative model can be defined as:

$$v_j \sim \text{Beta}(1, \alpha), \tag{42}$$

$$\pi_j = v_j \prod_{l=1}^{j-1} (1 - v_l), \tag{43}$$

$$(a_{jd}, b_{jd}, \lambda_{jd}) \sim G_0, \tag{44}$$

$$z_i \sim \text{Categorical}(\pi), \tag{45}$$

$$x_{id} \sim \text{LNB}(a_{z_i d}, b_{z_i d}, \lambda_{z_i d}) \tag{46}$$

Equation (42) specifies that the stick-breaking proportions v_j follow a Beta distribution with parameters 1 and α. The mixture weights π_j are then obtained using the formulation in (43). The component-specific parameters $(a_{jd}, b_{jd}, \lambda_{jd})$ are drawn from the base distribution G_0 as shown in (44). The cluster assignment for each observation is modeled by the categorical distribution in (45), and the likelihood of the data x_{id} is governed by an LNB distribution in (46), parameterized by the cluster-specific parameters. For computational tractability, we assume conditional independence across dimensions, allowing the multivariate density to be expressed as a product of univariate LNB densities:

$$p(\mathbf{x}_i \mid z_i, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) = \prod_{d=1}^{D} p(x_{id} \mid a_{z_i d}, b_{z_i d}, \lambda_{z_i d}) \tag{47}$$

The model with the joint distribution of all variables can be written as:

$$p(X \mid \boldsymbol{\Theta}) = \prod_{i=1}^{N} \sum_{j=1}^{\infty} \pi_j \prod_{d=1}^{D} p(x_{id} \mid a_{jd}, b_{jd}, \lambda_{jd}) \tag{48}$$

$$\begin{aligned} p(X, Z, V, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda} \mid \alpha) &= p(V \mid \alpha)\, p(\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})\, p(Z \mid V)\, p(X \mid Z, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) \\ &= \prod_{j=1}^{\infty} p(v_j \mid \alpha) \times \prod_{j=1}^{\infty} \prod_{d=1}^{D} p(a_{jd})\, p(b_{jd})\, p(\lambda_{jd}) \times \prod_{i=1}^{N} p(z_i \mid V)\, p(\mathbf{x}_i \mid z_i, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}). \end{aligned}$$

where:

$$p(v_j \mid \alpha) = \text{Beta}(v_j \mid 1, \alpha) \tag{49}$$

$$p(z_i = j \mid V) = \pi_j = v_j \prod_{l=1}^{j-1} (1 - v_l) \tag{50}$$

$$p(\mathbf{x}_i \mid z_i = j, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) = \prod_{d=1}^{D} p(x_{id} \mid a_{jd}, b_{jd}, \lambda_{jd}) \tag{51}$$

3.2 Variational Inference Framework

To make inference tractable and computationally efficient, we employ a truncated stick breaking approximation with a finite number of components K, where K is chosen to be sufficiently large.

This approach sets $v_K = 1$, effectively truncating the stick-breaking process after K components, while converging to the true Dirichlet process as $K \to \infty$. The truncated stick-breaking prior becomes $p(v_j|\alpha) = \text{Beta}(v_j|1, \alpha)$ for $j < K$ and $p(v_j|\alpha) = 1$ for $j = K$. Ensuring that $\sum_{j=1}^{K} \pi_j = 1$ and enabling us to work with a finite parameter set. We approximate the true posterior distribution $p(Z, V, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}|X)$ with a factorized variational distribution:

$$q(Z, V, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) = q(Z)q(V)\prod_{j=1}^{K}\prod_{d=1}^{D} q(a_{jd})q(b_{jd})q(\lambda_{jd}) \tag{52}$$

$$q(Z) = \prod_{i=1}^{N} q(z_i) = \prod_{i=1}^{N}\prod_{j=1}^{K} r_{ij}^{I(z_i=j)} \tag{53}$$

$$q(V) = \prod_{j=1}^{K-1} \text{Beta}(v_j|\gamma_{j1}, \gamma_{j2}) \tag{54}$$

Our variational inference framework approximates the true posterior distribution by optimizing a set of variational parameters. Specifically, we optimize three groups of parameters: the component assignment probabilities r_{ij} that determine the mixture component allocation for each observation, the Beta distribution parameters γ_{j1} and γ_{j2} that govern the stick-breaking construction of mixture weights, and the point estimates $\mu_{a_{jd}}$, $\mu_{b_{jd}}$, and $\mu_{\lambda_{jd}}$ that characterize the LNB parameters through approximations. These parameters are iteratively refined to minimize the KL divergence between the variational distribution and the true posterior.

3.2.1 Evidence Lower Bound

In our proposed Variational inference framework, we maximize the evidence lower bound (ELBO). A simplified ELBO formulation can be formulated as follows:

$$L = E_q[\log p(X, Z, V, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda}) - \log q(Z, V, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})] \tag{55}$$

which decomposes into the following terms:

$$L = \underbrace{E_q[\log p(X|Z, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})]}_{\text{Data likelihood}} + \underbrace{E_q[\log p(Z|V)]}_{\text{Assignment likelihood}} + \underbrace{E_q[\log p(V|\alpha)]}_{\text{Stick-breaking prior}} + \underbrace{E_q[\log p(\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})]}_{\text{Parameter prior}}$$
$$- \underbrace{E_q[\log q(Z)]}_{\text{Assignment entropy}} - \underbrace{E_q[\log q(V)]}_{\text{Stick-breaking entropy}} - \underbrace{E_q[\log q(\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})]}_{\text{Parameter entropy}}$$

The variational lower bound consists of several terms that together approximate the marginal likelihood of our model. We begin with the expected data log likelihood which is expressed as:

$$E_q[\log p(X|Z, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})] = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij} \times \sum_{d=1}^{D} \log p(x_{id}|\mu_{a_{jd}}, \mu_{b_{jd}}, \mu_{\lambda_{jd}})$$

When we substitute the log likelihood of the LNB distribution, this expands to:

$$E_q[\log p(X|Z, \mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})] = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij} \sum_{d=1}^{D} \Big[\mu_{a_{jd}} \log \mu_{\lambda_{jd}} + (\mu_{a_{jd}} - 1)\log x_{id} + (\mu_{b_{jd}} - 1)\log(1 - x_{id}) - \log B(\mu_{a_{jd}}, \mu_{b_{jd}}) - (\mu_{a_{jd}} + \mu_{b_{jd}})\log(1 - (1 - \mu_{\lambda_{jd}})x_{id}) \Big]$$

The cluster assignments in our mixture model are governed by the expected assignment log likelihood. This term captures how data points are probabilistically assigned to different components and is given by:

$$E_q[\log p(Z|V)] = \sum_{i=1}^{N}\sum_{j=1}^{K} r_{ij} E_q[\log \pi_j] \tag{56}$$

The expected log mixture weights are derived from the stick-breaking construction, which ensures that the mixture weights sum to one.

$$E_q[\log \pi_j] = E_q[\log v_j] + \sum_{l=1}^{j-1} E_q[\log(1 - v_l)] \tag{57}$$

For the Beta-distributed stick-breaking variables, the expected logarithms are computed using the digamma function $\psi(\cdot)$:

$$E_q[\log v_j] = \psi(\gamma_{j1}) - \psi(\gamma_{j1} + \gamma_{j2}) \tag{58}$$

$$E_q[\log(1 - v_j)] = \psi(\gamma_{j2}) - \psi(\gamma_{j1} + \gamma_{j2}) \tag{59}$$

The prior over the stick-breaking variables introduces a concentration parameter α that controls the tendency to use fewer or more mixture components. This expected stick-breaking prior is follows

$$E_q[\log p(V|\alpha)] = \sum_{j=1}^{K-1} [\log \alpha + (\alpha - 1)(\psi(\gamma_{j2}) - \psi(\gamma_{j1} + \gamma_{j2}))] \tag{60}$$

For the parameters of the LNB distribution, we incorporate Gamma priors to regularize the inference. Where s_{jd}, l_{jd}, u_{jd}, r_{jd}, v_{jd}, and g_{jd} are hyperparameters for the LNB distribution parameters. The expected parameter prior term is:

$$E_q[\log p(\mathbf{a}, \mathbf{b}, \boldsymbol{\lambda})] = \sum_{j=1}^{K} \sum_{d=1}^{D} \Big[\log p(a_{jd} = \mu_{a_{jd}} | s_{jd}, l_{jd}) + \log p(b_{jd} = \mu_{b_{jd}} | u_{jd}, r_{jd}) \tag{61}$$

$$+ \log p(\lambda_{jd} = \mu_{\lambda_{jd}} | v_{jd}, g_{jd}) \Big] \tag{62}$$

The variational inference framework introduces entropy terms that prevent the approximate posterior from collapsing to a point estimate. The assignment entropy represents the uncertainty in cluster assignments.

$$-E_q[\log q(Z)] = -\sum_{i=1}^{N} \sum_{j=1}^{K} r_{ij} \log r_{ij} \tag{63}$$

Similarly, the stick-breaking entropy quantifies the uncertainty in the mixture weights.

$$-E_q[\log q(V)] = \sum_{j=1}^{K-1} \Big[\log B(\gamma_{j1}, \gamma_{j2}) - (\gamma_{j1} - 1)\psi(\gamma_{j1}) - (\gamma_{j2} - 1)\psi(\gamma_{j2}) \Big.$$

$$\Big. + (\gamma_{j1} + \gamma_{j2} - 2)\psi(\gamma_{j1} + \gamma_{j2}) \Big] \tag{64}$$

By combining these terms, we obtain the complete variational lower bound that guides the optimization process. This bound becomes tighter as our approximate posterior better captures the true posterior distribution, leading to more accurate inference of the underlying mixture model. Finally, our inference procedure iteratively optimizes the variational lower bound by updating three groups of parameters in sequence. A representation of Variational Inference approach can be viewed in

Algorithm 2 Variational Inference Algorithm

Require: Data $X = \{\mathbf{x}_i\}_{i=1}^{N}$, truncation level K, concentration parameter α
Ensure: Variational parameters $\{r_{ij}\}$, $\{\gamma_{j1}, \gamma_{j2}\}$, $\{\mu_{a_{jd}}, \mu_{b_{jd}}, \mu_{\lambda_{jd}}\}$

Initialize variational parameters
repeat
  // Update component responsibilities (E-step)
  for $i = 1$ **to** N **do**
    for $j = 1$ **to** K **do**
      $\log r_{ij} \leftarrow E_q[\log \pi_j] + \sum_{d=1}^{D} \log p(x_{id} \mid \mu_{a_{jd}}, \mu_{b_{jd}}, \mu_{\lambda_{jd}})$
    end for
    $r_{ij} \leftarrow \frac{\exp(\log r_{ij})}{\sum_{j'=1}^{K} \exp(\log r_{ij'})}$ for all j ▷ Normalization
  end for
  // Update stick-breaking parameters (M-step)
  for $j = 1$ **to** K **do**
    $\gamma_{j1} \leftarrow 1 + \sum_{i=1}^{N} r_{ij}$
    $\gamma_{j2} \leftarrow \alpha + \sum_{i=1}^{N} \sum_{l=j+1}^{K} r_{il}$
  end for
  // Update Libby-Novick Beta parameters (M-step)
  for $j = 1$ **to** K **do**
    for $d = 1$ **to** D **do**
      $(\mu_{a_{jd}}, \mu_{b_{jd}}, \mu_{\lambda_{jd}}) \leftarrow \arg\min_{a,b,\lambda} \left\{ -\sum_{i=1}^{N} r_{ij} \log p(x_{id} \mid a, b, \lambda) \right\}$
    end for
  end for
  Calculate ELBO value for current iteration
until relative ELBO change below 10^{-6} for 3 consecutive iterations

Algorithm.2. This improves the objective function while maintaining the probabilistic constraints.

3.3 Experimental Setup & Results

Our proposed model was evaluated using three imaging datasets. The datasets were selected from the medical domain to demonstrate real world applicability in medical contexts which is often challenging. Our evaluation builds on a three stage preprocessing method. Firstly, feature extraction which is conducted via Scale-Invariant Feature Transform (SIFT) [70], Secondly, is the representation of extracted features through Bag of Visual Words methodology [71]. We utilize SIFT for feature extraction in our medical imaging analysis due to its scale and rotation invariance properties, which are critical for extracting consistent keypoints from variable medical images. Moreover, SIFT has the ability to identify distinctive local features serves as an ideal foundation for our Bag

of Visual Words representation, maintaining spatial information that is essential for clustering models. Finally, we performed feature scaling through unit-range normalization, mapping all values to the [0,1] interval. This scaling ensured the adherence to the LNB distribution assumption's on the bounded domain requirements. For comparative analysis, we benchmarked our approach against Gaussian Mixture Model (GMM) and Dirichlet Process Gaussian Mixture Model (DPGMM) [72]. This comparison showcases two key strengths of our approach. First, the GMM serves as a parametric baseline, which relies on Gaussian assumptions and a fixed component count. Second, the DPGMM's ability to infer its complexity via a stick-breaking prior emphasizes how our proposed variational framework not only preserves automatic model selection but also yields more accurate clustering in bounded domains.

To evaluate the performance of our proposed methodology and baseline models, we employed complementary clustering metrics. Firstly, the cluster Purity metric, which quantifies the proportion of correctly classified instances within each cluster. Moreover, the Fowlkes-Mallows Index (FMI) calculates the geometric mean of precision and recall, where higher values denote greater balance between clustering results and ground truth labels. Both of the aforementioned metrics are bounded between 0 and 1. Additionally, we also used the Normalized Mutual Information (NMI) which is also bounded between 0 and 1. For chance adjusted assessment, we utilized the Adjusted Rand Index (ARI), which lies within the [-1,1] interval, where positive values indicate meaningful cluster agreement.

3.3.1 Leukemia Detection

Leukemia is a blood cancer that affects the body's ability to produce normal blood cells and fight infections. Acute lymphoblastic leukemia (ALL) is a fast-growing type that primarily impacts lymphoid cells in the bone marrow. This condition requires prompt medical intervention and can affect both children and adults [73]. ALL progresses rapidly, with immature blasts demonstrating significantly impaired functionality compared to normal lymphocytes. In this study we use the Leukemia dataset proposed in [74] with 700 samples per class. Fig.3.2 showcases some samples from the dataset.

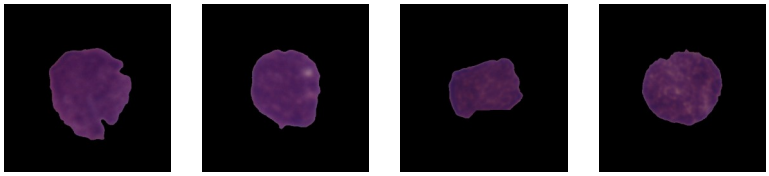

Figure 3.2: Representative samples of uninfected blood cells from Acute Lymphoblastic Leukemia (ALL) dataset.

Our experimental results, as shown in Table 3.1, demonstrate that the proposed VILNBMM model outperforms traditional clustering approaches. Notably, VILNBMM achieved the highest purity score of 0.864. Furthermore, it has also achieved and FMI score of 0.860, reflecting enhanced clustering accuracy.

Table 3.1: Comparison of Models on the Leukemia Dataset

Metric	Model		
	VILNBMM	GMM	DPGMM
Purity	0.864	0.836	0.853
FMI	0.860	0.841	0.853
ARI	0.529	0.451	0.462
NMI	0.435	0.379	0.347

3.3.2 Lung Cancer Diagnosis

Lung cancer remains a leading cause of cancer mortality worldwide, with adenocarcinoma and squamous cell carcinoma being the most common subtypes[67]. Our study utilized the lung cancer subset of the dataset proposed in [68], which contains histopathological images across the following three classes, lung adenocarcinoma, lung squamous cell carcinoma, and benign lung tissue. Each class consists of equal numbers of color histopathological images. In this study we use 500 samples per class. Fig. 4.2 showcases some samples from the dataset. Our analysis of the Lung Dataset in Table 3.2 demonstrates that the proposed VILNBMM model outperforms the benchmarking clustering approaches, with a purity score of 0.892.

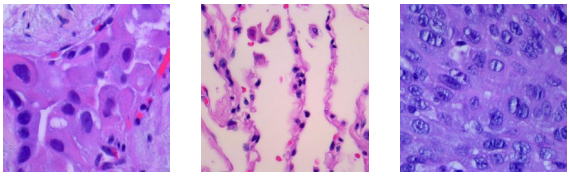

Figure 3.3: Representative histopathological images from the lung cancer dataset, spanning adenocarcinoma, benign tissue and squamous cell carcinoma.

Table 3.2: Comparison of Models on the Lung Dataset

Metric	Model		
	VILNBMM	GMM	DPGMM
Purity	0.892	0.857	0.846
FMI	0.878	0.845	0.839
ARI	0.682	0.635	0.488
NMI	0.654	0.596	0.501

3.3.3 Malaria Detection

Malaria represents a significant health threat requiring prompt and precise diagnosis for effective treatment. Traditional diagnosis relies on microscopic examination of blood smears, where clinicians identify infected cells based on visible differences compared to healthy cells [64]. Sample images are presented in Fig.3.4. This diagnostic approach presents challenges due to the need to analyze numerous cells with high accuracy. In our setup we have utilized 1,400 images that equally distrubuted between the two classes dataset that is obtained from [65].

Our evaluation on the Malaria Dataset, as presented in Table 3.3, showcases the enhanced performance of the VILNBMM model. The VILNBMM achieved higher purity and FMI scores, demonstrating cluster homogeneity and alignment with ground truth labels. The advantage in ARI and NMI metrics further validates VILNBMM's enhanced clustering accuracy and information correlation with reference classes.

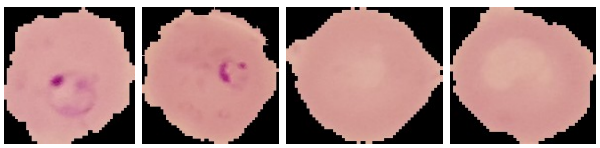

Figure 3.4: Representative images of blood cells from the malaria datase, infected and uninfected.

Our proposed VILNBMM outperforms GMM and DPGMM across the evaluation datasets, as

Table 3.3: Comparison of Models on Malaria Dataset

Metric	Model		
	VILNBMM	GMM	DPGMM
Purity	0.921	0.880	0.887
FMI	0.906	0.882	0.888
ARI	0.614	0.579	0.504
NMI	0.503	0.484	0.445

shown by the purity score visualization presented in Fig.3.5. The performance is particularly notable in the Malaria dataset, where VILNBMM achieves 4.1% improvement over GMM. Our approach demonstrates enhanced clustering capability across diverse medical imaging contexts, suggesting that the VILNBMM offers advantages for biomedical image clustering.

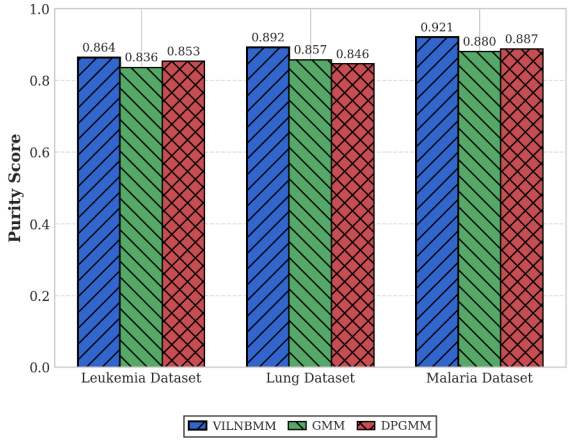

Figure 3.5: Purity score based Performance comparison of proposed VILNBMM against GMM and DPGMM baselines across datasets.

Chapter 4

Nonparametric Neural Variational Inference for McDonald's Beta Mixture Models with Feature Selection

In this chapter, we present NVI-IMBMM, a novel framework for infinite multivariate mixture modeling that combines the McDonald's Beta distribution with neural variational inference. Our approach addresses the fundamental challenges of automatic model selection, feature relevance determination, and enables posterior inference in clustering tasks. The methodology integrates the following key features. First, the Infinite multivariate McDonald's Beta mixture model with integrated feature selection (section 4.1). Secondly, the neural variational inference framework for posterior approximation in section 4.2.

Initially, let us consider a D-dimensional data point $\mathbf{x}_n$ represented by $(x_{n1}, \ldots, x_{nD})$ that follows the McBD. This distribution is characterized by shape parameters $\mathbf{a}_j = (a_{j1}, \ldots, a_{jD})$, $\mathbf{b}_j = (b_{j1}, \ldots, b_{jD})$, $\mathbf{p}_j = (p_{j1}, \ldots, p_{jD})$, and $\mathbf{q}_j = (q_{j1}, \ldots, q_{jD})$. It is worth noting that the parameters satisfy the constraints $a_{jd} > 0$, $b_{jd} > 0$, and $p_{jd} > 0$ for $d = 1, \ldots, D$. Moreover, $0 \leq x_{nd} \leq q_{jd}$ with $q_{jd} > 0$ for $d = 1, \ldots, D$. For the purpose of this study, we assume $q_{jd} = 1$ for all d, so as to constrain each x_{nd} to fall in the $[0, 1]$ range. Consequently, the joint density of the observation under this assumption is given by Equation (65), where the beta function is defined as

per Equation (66).

$$p(\mathbf{x}_n \mid \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \mathbf{p}_j, \mathbf{q}_j) = \prod_{d=1}^{D} \frac{p_{jd}\, x_{nd}^{a_{jd} p_{jd} - 1} \left(1 - x_{nd}^{p_{jd}}\right)^{b_{jd} - 1}}{B(a_{jd}, b_{jd})} \tag{65}$$

$$B(a_{jd}, b_{jd}) = \int_0^1 t^{a_{jd} - 1} (1 - t)^{b_{jd} - 1}\, dt = \frac{\Gamma(a_{jd})\,\Gamma(b_{jd})}{\Gamma(a_{jd} + b_{jd})} \tag{66}$$

By setting $q_{jd} = 1$ and $p_{jd} = 1$ in Equation (65), the McBD reduces to the ordinary Beta distribution. In our work, in order to constrain the support of the data between zero and one, we have made the assumption that q equals 1. A visualization of the McBD density function is shown in Figure 4.1.

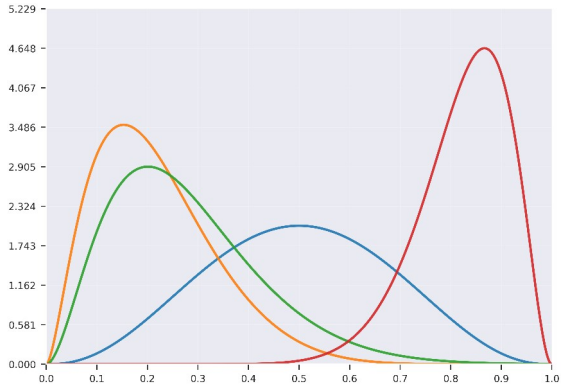

Figure 4.1: Comparison of McDonald's Beta probability density functions with different parameter configurations and degrees of flexibility. Parameters shown: symmetric distribution (blue), moderate left skew (orange), extreme left skew (green), and right skew (red).

4.1 Infinite McDonald's Beta Mixture with Feature Selection

We now derive the McDonald's Beta Mixture Model with feature selection (FMcDBMM), which generalizes the standard mixture model framework to simultaneously perform clustering and feature selection. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ denote a dataset of N independent and identically distributed observations, where each observation $\mathbf{x}_n = (x_{n1}, \ldots, x_{nD})$ is a D-dimensional vector with entries constrained to the unit interval $x_{nd} \in [0, 1]$. The mixture model with M components is defined by the density given in Equation (67), where $\boldsymbol{\theta}_j = (\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j, \mathbf{p}_j)$ denotes the parameters of the

McBD for component j, and the mixture weights $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ satisfy the standard constraints in Equation 68

$$p(X \mid \Theta) = \prod_{n=1}^{N} \sum_{j=1}^{M} \pi_j \, p(\mathbf{x}_n \mid \boldsymbol{\theta}_j) = \prod_{n=1}^{N} \sum_{j=1}^{M} \pi_j \prod_{d=1}^{D} \frac{p_{jd} \, x_{nd}^{a_{jd}\, p_{jd} - 1} \, (1 - x_{nd})^{b_{jd} - 1}}{B(a_{jd}, b_{jd})} \tag{67}$$

$$\sum_{j=1}^{M} \pi_j = 1, \qquad \pi_j \geq 0 \quad \text{for all } j \tag{68}$$

To enable automatic feature selection, we extend this model by introducing binary indicator variables $\delta_{jd} \in \{0, 1\}$, which signals whether feature d is considered to be relevant for component j. The component specific likelihood for observation $\mathbf{x}_n$ assigned to cluster j becomes as shown in Equation (69). where $f(x_{nd} \mid \theta_{jd})$ is the McBD with parameters $\theta_{jd} = (a_{jd}, b_{jd}, p_{jd})$, and $g(x_{nd} \mid \xi_d)$ is a shared background density with parameters $\xi_d = (a_d, b_d, p_d)$, used when feature d is deemed irrelevant to cluster j. The binary indicators are modeled using independent Bernoulli priors as defined in Equation (70), where $\rho_{jd} = P(\delta_{jd} = 1)$ encodes the saliency of feature d for component j. Marginalizing over the feature indicators yields the likelihood for component j as shown in Equation (71). which provides a soft weighting between cluster-specific and background treatment for each feature.

$$p(\mathbf{x}_n \mid z_n = j, \theta, \xi, \delta) = \prod_{d=1}^{D} [f(x_{nd} \mid \theta_{jd})]^{\delta_{jd}} \, [g(x_{nd} \mid \xi_d)]^{1 - \delta_{jd}} \tag{69}$$

$$p(\delta_{jd} \mid \rho_{jd}) = \rho_{jd}^{\delta_{jd}} \, (1 - \rho_{jd})^{1 - \delta_{jd}} \tag{70}$$

$$p(\mathbf{x}_n \mid z_n = j, \theta, \xi, \rho) = \prod_{d=1}^{D} \rho_{jd} \, f(x_{nd} \mid \theta_{jd}) + (1 - \rho_{jd}) \, g(x_{nd} \mid \xi_d) \tag{71}$$

More specifically stated, when $\rho_{jd} \approx 1$, feature d is highly informative for cluster j; conversely, when $\rho_{jd} \approx 0$, the feature is uninformative and modeled using the background distribution. The complete data likelihood under this feature-selective mixture is given by Equation (72), where π, θ, ξ, and ρ fully characterize the model.

$$p(X \mid \pi, \theta, \xi, \rho) = \prod_{n=1}^{N} \sum_{j=1}^{M} \pi_j \, p(\mathbf{x}_n \mid z_n = j, \theta, \xi, \rho) \tag{72}$$

While the finite mixture model provides a solid foundation, selecting the appropriate number of components M remains a significant challenge in practical applications. To address this limitation, we extend our framework to a nonparametric Bayesian setting using the DP. This is to enable the model to automatically determine the optimal number of clusters from the data while maintaining the feature selection properties. The DP provides an approach to infinite mixture modeling through the constructive definition via the stick-breaking process. We place a DP prior over the mixture components with concentration parameter $\alpha_0 > 0$ and base distribution G_0 that governs the generation of component specific parameters. The stick-breaking construction proceeds by drawing an infinite sequence of breaking proportions $\beta_j \sim \text{Beta}(1, \alpha_0)$ for $j = 1, 2, \ldots$, which define the mixture weights through the relationships shown in Equation (78). This construction ensures that $\sum_{j=1}^{\infty} \pi_j = 1$, providing a probability distribution over an infinite number of components. The concentration parameter α_0 controls the expected number of active components, with larger values encouraging more clusters. For each component j, we independently sample its parameters from the base distribution G_0, which factorizes as $G_0 = H_\theta \times H_\rho \times H_\xi$, where H_θ governs the McDonald's Beta parameters, H_ρ controls the feature saliency probabilities, and H_ξ determines the background distribution parameters. Specifically, we specify the prior distributions as given in Equation (74).

$$\pi_1 = \beta_1, \quad \pi_j = \beta_j \prod_{\ell=1}^{j-1} (1 - \beta_\ell) \quad \text{for } j \geq 2 \tag{73}$$

$$\begin{aligned} a_{jd}, b_{jd}, \rho_{jd} &\sim \text{Gamma}(\alpha_a, \beta_a) \times \text{Gamma}(\alpha_b, \beta_b) \times \text{Gamma}(\alpha_\rho, \beta_\rho) \\ \rho_{jd} &\sim \text{Beta}(\kappa_1, \kappa_2) \end{aligned} \tag{74}$$

$$\begin{aligned} a_d, b_d, \rho_d &\sim \text{Gamma}(\alpha_a, \beta_a) \times \text{Gamma}(\alpha_b, \beta_b) \\ &\times \text{Gamma}(\alpha_\rho, \beta_\rho) \end{aligned} \tag{75}$$

The choice of $\kappa_1 < \kappa_2$ in the Beta prior for ρ_{jd} encourages sparse feature usage by biasing the saliency probabilities toward smaller values, which promotes automatic feature selection. In practice, we employ a truncated stick-breaking approximation by setting a sufficiently large truncation level K, effectively capping the number of components while maintaining the essential properties

of the infinite model.

4.2 Neural Variational Inference Framework

The inference challenge in the infinite McDonald's Beta mixture model stems from the intractability of the posterior distribution. On the contrary to conjugate models where closed form updates exist, the McBD has additional parameters. This structure prevents the computation for the posterior analytically, it also necessitates approximate inference methods. We address this challenge through a neural variational inference framework that combines deep learning with Bayesian inference. In our framework, instead of maintaining separate variational parameters for each data point, we employ neural networks as function approximators that directly map observations to posterior distributions. The cornerstone of our inference framework lies in the variational family. We approximate the intractable true posterior $p(Z, \pi, \theta, \xi, \rho \mid X)$through a structured mean-field factorization that preserves essential dependencies while enabling tractable computation as shown in Equation 76. While this factorization assumes independence between different parameter groups, each of the factors maintain internal structure that captures posterior patterns.

$$q(Z, \pi, \theta, \xi, \rho) = q(Z)\, q(\pi)\, q(\theta)\, q(\rho)\, q(\xi) \tag{76}$$

Initially, for the cluster assignments we use categorical distributions that naturally represent observations to mixture components, as formalized in Equation 77. where ϕ_{nj} represents the posterior probability that observation n belongs to cluster j. These soft assignments actually enable gradient based optimization while maintaining interpretability as probability distributions.

$$q(z_n = j) = \phi_{nj}, \quad \sum_{j=1}^{K} \phi_{nj} = 1 \tag{77}$$

The stick-breaking construction is fundamental to the DP. It is integrated via the stick-breaking proportions through Beta distributions, this is to mirror the prior structure, as shown in Equation 78. This choice ensures conjugacy with the Beta prior, which subsequently facilitates analytical KL

divergence computation while maintaining the unit interval constraint to the stick-breaking itself.

$$q(\beta_j) = \text{Beta}(\gamma_{j1}, \gamma_{j2}), \quad j = 1, \ldots, K-1 \tag{78}$$

The McBD parameters pose challenges due to their positivity constraints. For the aforementioned reason we address this through LogNormal variational distributions that naturally enforce positivity while providing flexibility, as specified in Equation 79. The motivations for using the LogNormal family are through some advantages that it offers, such as it naturally handles the heavy-tailed distributions, and provides straightforward reparameterization for gradient estimation. The feature saliency indicators are the key parts to the automatic feature selection mechanism in our proposed model. These saliency indicators are modeled through Beta distributions that capture the uncertainty in feature relevance, as expressed in Equation 80. This parameterization allows interpolation between relevant ($\rho_{jd} \approx 1$) and irrelevant ($\rho_{jd} \approx 0$) features while maintaining differentiability which is needed for the gradient based optimization. The background distribution parameters are shared across all irrelevant features, and follow a similar LogNormal parameterization given by Equation 81.

$$\begin{aligned} q(a_{jd}, b_{jd}, \rho_{jd}) = \; & \text{LogNormal}(\mu_{a_{jd}}, \sigma^2_{a_{jd}}) \\ & \times \text{LogNormal}(\mu_{b_{jd}}, \sigma^2_{b_{jd}}) \times \text{LogNormal}(\mu_{\rho_{jd}}, \sigma^2_{\rho_{jd}}) \end{aligned} \tag{79}$$

$$q(\rho_{jd}) = \text{Beta}(\lambda^{(1)}_{jd}, \lambda^{(2)}_{jd}) \tag{80}$$

$$q(a_d, b_d, \rho_d) = \text{LogNormal}(\mu_{a_d}, \sigma^2_{a_d}) \times \text{LogNormal}(\mu_{b_d}, \sigma^2_{b_d}) \times \text{LogNormal}(\mu_{\rho_d}, \sigma^2_{\rho_d}) \tag{81}$$

The transformation from raw observations to variational parameters requires an architectural design that aims to balance expressiveness and generalization. Our neural amortization method employs specialized networks for the different parameter groups. Each network is tailored to meet any of the specific inferential requirements. The encoder network serves as the primary inference mechanism which maps observations to cluster assignment probabilities through a multi-layer perceptron architecture (MLP), as shown in Equation 82. The network f_{enc} comprises 3 hidden layers with ReLU activations. The softmax output layer produces valid probability distributions over clusters,

this helps in maintaining the constraint of $\sum_{j=1}^{K} \phi_{nj} = 1$ automatically.

$$\boldsymbol{\phi}_n = \text{softmax}(f_{\text{enc}}(\mathbf{x}_n; \boldsymbol{\psi})) \tag{82}$$

The depth and width of the architecture are adjusted acording to the complexity of the data. Typically shallow networks with wider layers are sufficient for well separated clusters, while deeper architectures are necessary for capturing subtle patterns. We have also used batch normalization between layers which helps in stabilizing the training, particularly important given the diverse scales of features in real world medical applications. The Parameter generation for MCBD distributions and feature saliency uses a different strategy, here we leverage learnable embeddings to enable parameter sharing across clusters, as shown in Equations 83 and 84. where $\mathbf{e}_j \in R^E$ represents a learnable embedding for cluster j.

$$[\boldsymbol{\mu}_a, \boldsymbol{\sigma}_a, \boldsymbol{\mu}_b, \boldsymbol{\sigma}_b, \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p] = f_{\text{param}}(\mathbf{e}_j; \boldsymbol{\omega}) \tag{83}$$

$$[\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}] = f_{\text{saliency}}(\mathbf{e}_j; \boldsymbol{\nu}) \tag{84}$$

Building upon the principles of VI, where our goal is essentially to optimize a function, this is where we maximize a lower bound on the log marginal likelihood. The Evidence Lower Bound (ELBO) decomposes into interpretable components that reflect different aspects of the model, as expressed in Equation 85. The expected joint probability decomposes as per our model's generative process, while the entropy term encourages exploration of the posterior. Expanding these expectations gives the detailed form in Equation 86.

$$L(\boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{\nu}) = E_q[\log p(X, Z, \pi, \theta, \xi, \rho)] - E_q[\log q(Z, \pi, \theta, \xi, \rho)] \tag{85}$$

$$L = \sum_{n=1}^{N}\sum_{j=1}^{K} \phi_{nj} E_{q(\boldsymbol{\theta},\boldsymbol{\xi},\boldsymbol{\rho})} [\log p(\mathbf{x}_n \mid z_n = j, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\rho})] + \sum_{n=1}^{N}\sum_{j=1}^{K} \phi_{nj} E_{q(\boldsymbol{\pi})} [\log \pi_j] + \sum_{n=1}^{N} H[q(z_n)]$$
$$- \mathrm{KL}[q(\boldsymbol{\beta}) \| p(\boldsymbol{\beta} \mid \alpha_0)] - \sum_{j,d} \mathrm{KL}[q(\theta_{jd}) \| p(\theta_{jd})] - \sum_{j,d} \mathrm{KL}[q(\rho_{jd}) \| p(\rho_{jd})] - \sum_{d} \mathrm{KL}[q(\xi_d) \| p(\xi_d)] \tag{86}$$

The expected likelihood measures how well the model explains the observed data under the current posterior. The prior term on assignments incorporates the stick-breaking structure. The entropy of assignments prevents premature convergence to hard assignments, maintaining uncertainty throughout training. The KL divergence terms act as adaptive regularizers, preventing the variational distributions from deviating excessively or too far from their priors. In our model we aim to optimize the ELBO and this optimization involves treatment for both discrete and continuous latent variables. For continuous parameters, we employ the reparameterization trick, expressing samples as deterministic transformations of noise variables, as shown in Equation 87. This reparameterization enables us to backpropagate through the sampling operation itself, which typically gives low-variance gradient estimates, this is essential for the optimization's stability. The discrete cluster assignments present a different challenge, as the non-differentiability of sampling from categorical distributions prevents direct gradient computation. We address this through a differentiable cluster assignment mechanism which maintains gradient flow while approximating discrete categorical sampling similar to [75] and is formalized in Equation 88.

$$\theta = \exp(\mu + \sigma \cdot \epsilon), \quad \epsilon \sim N(0, 1) \tag{87}$$

$$z_{nj} = \frac{\exp((\log \phi_{nj} + g_{nj})/\tau)}{\sum_{k=1}^{K} \exp((\log \phi_{nk} + g_{nk})/\tau)} \tag{88}$$

where $g_{nj} \sim (0, 1)$ introduces stochasticity and the temperature parameter τ controls the approximation quality. As $\tau \to 0$, the distribution approaches a discrete categorical, where greater values produce smoother and more uniform distributions. The temperature is gradually decreased during training, this enables us to begin with values that encourages exploration. Generally, stochastic optimization proceeds through mini-batch gradient ascent where each iteration processes a subset

B of the data. The batch objective is given by Equation 89. This decomposition properly weights the contribution of batch specific terms such as likelihood and assignment entropy, and the global terms (KL divergences parameters). Our proposed infinite mixture model naturally adapts to data complexity through the automatic selection of the number of clusters. Moreover, during training we monitor the cluster utilization to identify and potentially prune inactive components, using the utilization metric defined in Equation 90.

$$L_{\text{batch}} = \frac{N}{|B|} \sum_{n \in B} L_n + \frac{1}{N} L_{\text{global}} \tag{89}$$

$$\text{Utilization}_j = \frac{1}{N} \sum_{n=1}^{N} \phi_{nj} \tag{90}$$

To consider a component inactive, we measure the utilization; if it falls below the threshold of $\epsilon_{\text{prune}} = 0.01$, then it is deemed to be inactive, reflecting insufficient data support for its continued inclusion. This adaptive pruning operationalizes the DP's property where the "rich get richer". In the aforementioned property, the popular clusters attract more observations while the unused components naturally fade away. We cap the truncation level K, which serves as an upper bound on model complexity rather than a fixed constraint. While theoretically the DP can handle an infinite number of clusters, this is unfeasible computationally. Practically, we set $K = 50$, which allows the model to have sufficient flexibility. We employ a multi-criteria convergence assessment. Training is terminated when the ELBO improvement falls below $\epsilon_{\text{conv}} = 10^{-4}$ for consecutive epochs. To prevent overfitting, we monitor three complementary indicators. First, the stability of cluster assignment across iterations, the gradient norm magnitudes to detect vanishing or exploding gradients, and the effective number of active components to track model complexity. Algorithm 3 shows the training procedure. Throughout training, we employ temperature annealing with an exponential decay schedule $\tau_{t+1} = \max(\tau_t \times 0.999, \tau_{\min})$, this progressively sharpens the cluster assignments by transitioning from exploratory soft assignments in early epochs to better allocations as training progresses.

Algorithm 3 Neural Variational Inference for NVI-IMBMM

Input: Dataset X, truncation level K, learning rate η, batch size B
Initialize: Neural network parameters $\boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{\nu}$, temperature $\tau = 1.0$
while not converged **do**
    Sample mini-batch $B \subset X$ of size B
    for $\mathbf{x}_n \in B$ **do**
        $\boldsymbol{\phi}_n \leftarrow \text{softmax}(f_{\text{enc}}(\mathbf{x}_n; \boldsymbol{\psi}))$
    end for
    Sample $\boldsymbol{\theta}, \boldsymbol{\rho}, \boldsymbol{\xi}$ using reparameterization
    Sample soft assignments $\tilde{\mathbf{z}}$
    $L_{\text{batch}} \leftarrow$ Compute ELBO on mini-batch
    $\boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{\nu} \leftarrow$ Update on L_{batch}
    $\tau \leftarrow \max(\tau \times 0.999, 0.5)$
end while
Return: Trained parameters $\boldsymbol{\psi}, \boldsymbol{\omega}, \boldsymbol{\nu}$

4.3 Experimental Setup and Results

We evaluated our proposed NVI-IMBMM model using three medical imaging datasets. Each selected specifically to demonstrate real-world applicability in challenging medical contexts. Our evaluation framework employs a three stage preprocessing pipeline. First, we extract features using the Scale-Invariant Feature Transform (SIFT) [70], this is to leverage its scale and rotation invariance properties to obtain consistent keypoints from variable medical images. Second, we represent the extracted features through the Bag of Visual Words methodology [71], which preserves essential information for clustering models while maintaining SIFT's distinctive local feature characteristics. Finally, we apply normalization to scale all feature values to the $[0, 1]$ interval, ensuring compliance with the bounded domain requirements. It is also worth clarifying that we adopted SIFT + BoVW due to its rotation and scale invariance and interpretability which are important for our clustering tasks. Moreover, BoVW produces bounded [0,1] vectors that are well suited for the bounded domain modeling.

For comparative analysis, and due to the probabilistic nature of out proposed model, we benchmarked our approach against two established probabilistic baselines. Firsly, the Gaussian Mixture Model (GMM) and the DP Gaussian Mixture Model (DPGMM) [72]. This comparison highlights two key advantages of our method. The GMM serves as a parametric baseline that relies on Gaussian assumptions with a fixed number of components, while the DPGMM demonstrates automatic

complexity inference through its stick-breaking prior.

To assess the performance of our methodology and baseline models, we employed a comprehensive set of complementary clustering metrics. We utilized cluster Purity which is computed as Purity $= \frac{1}{N} \sum_{i=1}^{k} \max_j |C_i \cap T_j|$, where N is the total number of data points, C_i represents the i-th cluster, and T_j denotes the j-th true class. The Fowlkes-Mallows Index (FMI) calculates the geometric mean of precision and recall as FMI $= \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}$, where TP, FP, and FN represent true positives, false positives, and false negatives respectively. For the FMI higher values indicate better balance between clustering results and ground truth labels. Additionally, we employed Normalized Mutual Information (NMI) which is defined as NMI $= \frac{I(C;T)}{\sqrt{H(C)H(T)}}$ where $I(C; T)$ is the mutual information between clusters C and true labels T, and $H(\cdot)$ denotes entropy. All three metrics are bounded within $[0, 1]$ with values closer to 1 indicating superior performance. For chance adjusted assessment, we incorporated the Adjusted Rand Index (ARI) and it is calculated as ARI $= \frac{RI - E[RI]}{\max(RI) - E[RI]}$ where RI is the Rand Index and $E[RI]$ is its expected value under random clustering. The ARI ranges from -1 to 1, where positive values signify meaningful cluster agreement beyond random chance.

4.3.1 Lung Cancer Classification

Lung cancer remains one of the leading causes of cancer related mortality worldwide [67], with adenocarcinoma and squamous cell carcinoma representing prevalent histological subtypes of the disease. Early and accurate classification of these subtypes is crucial for treatment planning and patient prognosis. We employed the lung cancer portion of the comprehensive histopathological dataset developed in [68], which comprises high-resolution microscopic tissue images systematically categorized into three distinct diagnostic groups. These groups are adenocarcinoma, squamous cell carcinoma, and healthy lung tissue. The dataset maintains class balance with identical quantities of colored histopathological images for each category. For our experimental validation, we analyzed 1000 images from each class, totaling 3000 tissue samples. Representative samples from each tissue type are illustrated in Figure 4.2.

Our experimental evaluation on the lung cancer dataset demonstrates that the proposed NVI-IMBMM model achieves superior performance compared to existing baseline clustering methodologies. As presented in Table 4.1, our approach achieved improvements. Our model achieved a Purity score of 0.935 (representing 7.8% and 8.9% improvements over GMM and DPGMM respectively), a Fowlkes-Mallows Index of 0.876, and most notably, an Adjusted Rand Index of 0.814, substantially outperforming both GMM (0.635) and DPGMM (0.488). The Normalized Mutual Information score of 0.786 further validates the effectiveness of our method in capturing meaningful relationships between image features and diagnostic categories. The most discriminative features identified by our proposed model through automated feature selection are visualized in Figure 4.3.

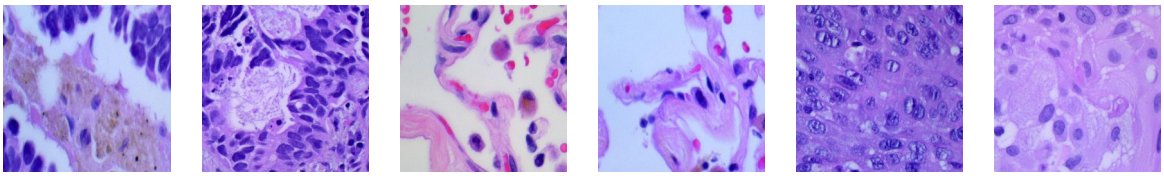

Figure 4.2: Representative samples from lung cancer dataset showing different tissue types: the first two images show adenocarcinoma, followed by two images of healthy lung tissue, and the final two images display squamous cell carcinoma.

Table 4.1: Clustering Performance Comparison on Lung Cancer Dataset

Metric	Model		
	GMM	DPGMM	NVI-IMBMM
Purity	0.857	0.846	0.935
FMI	0.845	0.839	0.876
ARI	0.635	0.488	0.814
NMI	0.596	0.501	0.786

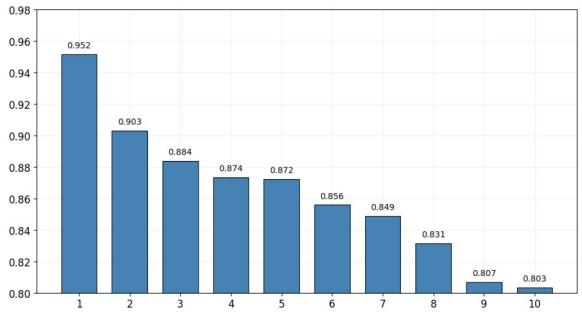

Figure 4.3: Top 10 most discriminative features identified by NVI-IMBMM for lung cancer, ranked by feature importance scores.

4.3.2 Skin Cancer detection

Skin cancer represents one of the most prevalent malignancies globally, with early detection being paramount for successful treatment outcomes and patient survival rates [76]. We utilized the skin cancer dataset from [77], which is a collection of images that provides standardized samples for malignancy detection. This dataset encompasses two clases which are malignant and benign skin cancer. For our experiments, we focused on the binary classification problem of distinguishing malignant from benign lesions, selecting 1000 images from each category to create a balanced dataset of 2000 samples. Figure 4.4 showcases samples from each class of the dataset.

Our experimental evaluation on the skin cancer dataset demonstrates the performance of the proposed NVI-IMBMM model when compared to conventional clustering approaches. As presented in Table 4.2, the NVI-IMBMM model attained a Purity score of 0.926, representing improvements of 9.8% and 8.3% over GMM and DPGMM respectively. The Fowlkes-Mallows Index reached 0.862, while the Adjusted Rand Index achieved 0.725, which outperforms both GMM (0.470) and DPGMM (0.498). Most notably, the Normalized Mutual Information score of 0.626 demonstrates the model's effectiveness in capturing meaningful discriminative patterns between malignant and benign lesion characteristics. The automated feature selection component of our approach identified the most diagnostically relevant features, as visualized in Figure 4.5, highlighting the model's ability to focus on morphological and textural attributes that distinguish malignant from benign skin lesions.

Table 4.2: Clustering Performance Comparison on Skin Cancer Dataset

Metric	Model		
	GMM	DPGMM	NVI-IMBMM
Purity	0.843	0.855	0.926
FMI	0.735	0.749	0.862
ARI	0.470	0.498	0.725
NMI	0.377	0.401	0.626

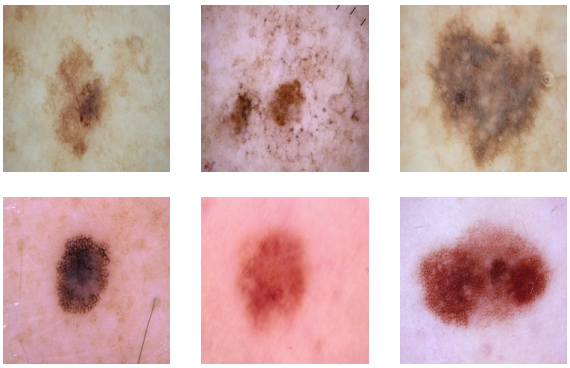

Figure 4.4: Representative samples from skin cancer dataset with malignant and benign tissue samples.

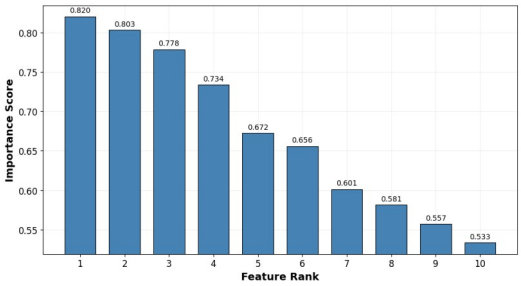

Figure 4.5: Top 10 most discriminative features identified by NVI-IMBMM for skin cancer detection.

4.3.3 Acute Lymphoblastic Leukemia Detection

Acute lymphoblastic leukemia (ALL) represents an aggressive and rapidly progressing form of blood malignancy that fundamentally disrupts the body's capacity to produce healthy blood cells and mount effective immune responses against infections [78]. The disease can manifest across all age demographics, from pediatric to adult populations, with particularly severe implications for children [73]. For this comprehensive investigation, we utilized the Acute Lymphoblastic Leukemia (ALL) detection dataset used in [79]. For the purpose of our study we used 500 images per class belonging to two categories, totaling 1000 cellular peripheral blood smear (PBS) images. The first class being the benign case and the second is (Pro-B ALL) which is a malignant subtype. The dataset provides representation of both healthy and leukemic cell populations, enabling evaluation of clustering performance. Representative cellular examples from both categories are illustrated in Figure 4.6. Our NVI-IMBMM model achieved excellent performance on the leukemia detection task, outperforming both baseline methods. The experimental results, presented in Table 4.3, demonstrate that our approach achieved a Purity score of 0.904, representing improvements of 7.3% over GMM (0.831) and 2.0% over DPGMM (0.884), with DPGMM showing competitive performance on this particular dataset. The Fowlkes-Mallows Index reached 0.826, indicating improved clustering quality compared to both GMM (0.726) and DPGMM (0.796). Our method achieved an Adjusted Rand Index of 0.651, outperforming GMM (0.438) by 21.3% and surpassing DPGMM (0.584) by 6.7%. The feature analysis revealed that our model successfully identified critical morphological characteristics as visualized in Figure 4.7.

Table 4.3: Clustering Performance Comparison on Acute Lymphoblastic Leukemia Dataset

Metric	Model		
	GMM	DPGMM	NVI-IMBMM
Purity	0.831	0.884	0.904
FMI	0.726	0.796	0.826
ARI	0.438	0.584	0.651
NMI	0.389	0.545	0.550

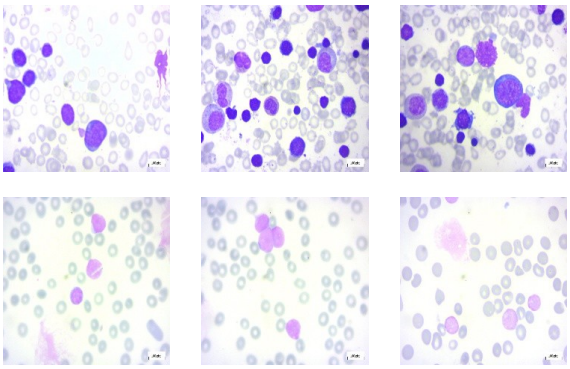

Figure 4.6: Representative samples from Acute Lymphoblastic Leukemia (ALL) dataset: normal lymphocytes (top row) showing typical mature cell morphology with well-defined nuclear-cytoplasmic boundaries, and leukemic blast cells (bottom row) displaying characteristic immature features including enlarged nuclei and reduced cytoplasm.

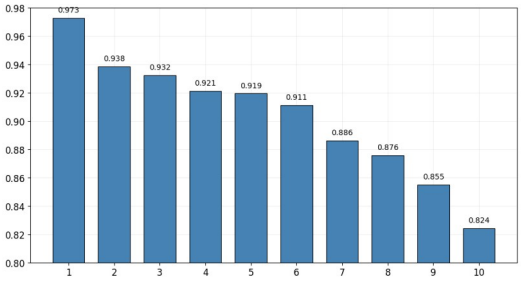

Figure 4.7: Top 10 most discriminative features identified by NVI-IMBMM for acute lymphoblastic leukemia detection.

Chapter 5

Conclusion

In this work, we aimed to address a fundamental problem of clustering bounded and asymmetric data in medical imaging applications. This is done through the development of three models that are based on flexible distributions within mixture modeling frameworks. The first approach presented a Bayesian Libby-Novick Beta mixture model that integrates automatic feature selection within the clustering process. The second contribution introduced the Variational Infinite Libby-Novick Beta Mixture Model (VILNBMM), which addresses the challenge of determining the optimal number of clusters in a bounded data setting. The third approach, NVI-IMBMM, presented a neural variational inference framework for infinite McDonald's Beta mixture model with embedded feature selection. The developed models use flexible Beta generalized distributions to capture the skewness and kurtosis that are present in bounded data. Additionally, the use of the Dirichlet processes enables automatic inference of cluster numbers. Furthermore, the adoption of variational inference across these frameworks transforms the intractability problem into optimization tasks. This was done through replacing sampling-based methods with optimization. Collectively, the proposed approaches demonstrate better modeling capabilities for bounded data compared to the traditional Gaussian assumptions within mixture modeling frameworks. Possible future directions can focus on the use of hybrid variational inference frameworks and can aim to use the discussed distributions within deep learning frameworks.

Bibliography

[1] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263–286, 2017.

[2] I. Johnstone and D. Titterington, "Statistical challenges of high-dimensional data," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 367, pp. 4237–53, 11 2009.

[3] A. E. Ezugwu, A. M. Ikotun, O. O. Oyelade, L. Abualigah, J. O. Agushaka, C. I. Eke, and A. A. Akinyelu, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022.

[4] A. Saxena, M. Prasad, A. Gupta, N. Bharill, O. P. Patel, A. Tiwari, M. J. Er, W. Ding, and C.-T. Lin, "A review of clustering techniques and developments," *Neurocomputing*, vol. 267, pp. 664–681, 2017.

[5] Z. Manbari, F. AkhlaghianTab, and C. Salavati, "Hybrid fast unsupervised feature selection for high-dimensional data," *Expert Systems with Applications*, vol. 124, pp. 97–118, 2019.

[6] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014, 40th-year commemorative issue.

[7] I. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, 03 2021.

[8] T. Elguebaly and N. Bouguila, "Simultaneous high-dimensional clustering and feature selection using asymmetric gaussian mixture models," *Image and Vision Computing*, vol. 34, pp. 27–41, 2015.

[9] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, p. 845–889, Dec. 2004.

[10] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[11] N. Bouguila and W. Fan, *Mixture Models and Applications*, 1st ed. Springer Publishing Company, Incorporated, 2019.

[12] G. J. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. Wiley, 2000.

[13] S. Boutemedjet, N. Bouguila, and D. Ziou, "A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1429–1443, 2009.

[14] M. Azam and N. Bouguila, "Multivariate bounded support asymmetric generalized gaussian mixture model with model selection using minimum message length," *Expert Systems with Applications*, vol. 204, p. 117516, 2022.

[15] L. Scrucca, C. Fraley, T. Murphy, and R. E., *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*, 03 2023.

[16] A. Algumaei, M. Azam, F. Najar, and N. Bouguila, "Bounded multivariate generalized gaussian mixture model using ica and iva," *Pattern Anal. Appl.*, vol. 26, no. 3, p. 1223–1252, Feb. 2023. [Online]. Available: https://doi.org/10.1007/s10044-023-01148-w

[17] A. Algumaei, M. Azam, M. Amayri, and N. Bouguila, "Adaptive constrained icabmggmm: Application to ecg blind source separation," in *Neural Information Processing*, 2025, pp. 107–123.

[18] N. Nasios and A. Bors, "Variational learning for gaussian mixture models," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 849–862, 2006.

[19] S. Chander and P. Vijaya, "3 - unsupervised learning methods for data clustering," in *Artificial Intelligence in Data Mining*, D. Binu and B. Rajakumar, Eds. Academic Press, 2021, pp. 41–64.

[20] K. Chen, T. van Laarhoven, and E. Marchiori, "Gaussian processes with skewed laplace spectral mixture kernels for long-term forecasting," *Mach. Learn.*, vol. 110, no. 8, p. 2213–2238, Aug. 2021.

[21] X. Shi, Y. Li, and Q. Zhao, "Flexible hierarchical gaussian mixture model for high-resolution remote sensing image segmentation," *Remote Sensing*, vol. 12, no. 7, 2020.

[22] E. Limpert and W. A. Stahel, "Problems with using the normal distribution – and ways to improve quality and efficiency of data analysis," *PLoS ONE*, vol. 6, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:18220560

[23] Z. Ma and A. Leijon, "Beta mixture models and the application to image classification," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 2045–2048.

[24] J. B. McDonald and Y. J. Xu, "A generalization of the beta distribution with applications," *Journal of Econometrics*, vol. 66, no. 1, pp. 133–152, 1995.

[25] Y. Ji, C. Wu, P. Liu, J. Wang, and K. R. Coombes, "Applications of beta-mixture models in bioinformatics," *Bioinformatics*, vol. 21, no. 9, p. 2118–2122, May 2005.

[26] S. Sinharay, "Continuous probability distributions," in *International Encyclopedia of Education (Third Edition)*, third edition ed., P. Peterson, E. Baker, and B. McGaw, Eds. Oxford: Elsevier, 2010, pp. 98–102.

[27] N. Bouguila, D. Ziou, and E. Monga, "Practical bayesian estimation of a finite beta mixture through gibbs sampling and its applications," *Statistics and Computing*, vol. 16, pp. 215–225, 06 2006.

[28] N. Manouchehri, M. Kalra, and N. Bouguila, "Online variational inference on finite multivariate beta mixture models for medical applications," *IET Image Processing*, vol. 15, no. 9, pp. 1869–1882, 2021. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/ipr2.12154

[29] K. Ketabchi, N. Manouchehri, and N. Bouguila, "Fully bayesian libby-novick beta mixture model with feature selection," in *2022 IEEE International Conference on Industrial Technology (ICIT)*, 2022, pp. 1–6.

[30] N. Samiee, N. Manouchehri, and N. Bouguila, "Finite libby-novick beta mixture model: An mml-based approach," in *Intelligent Information and Database Systems*, N. T. Nguyen, S. Boonsang, H. Fujita, B. Hnatkowska, T.-P. Hong, K. Pasupa, and A. Selamat, Eds. Singapore: Springer Nature Singapore, 2023, pp. 371–383.

[31] G. M. Cordeiro, L. H. de Santana, E. M. M. Ortega, and R. R. Pescim, "A new family of distributions: Libby-novick beta," *International Journal of Statistics and Probability*, vol. 3, p. 63, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:73641567

[32] N. Samiee, N. Manouchehri, and N. Bouguila, "A nonparametric bayesian framework for multivariate Libby-Novick beta mixture models," in *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2024, pp. 751–756.

[33] O. Sghaier, M. Amayri, and N. Bouguila, "Data clustering with libby-novick beta-liouville mixture models: A minimum message length approach," in *Proceedings of the 2024 9th International Conference on Intelligent Information Technology*, ser. ICIIT '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 314–321.

[34] D. Forouzanfar, N. Manouchehri, and N. Bouguila, "A fully bayesian inference approach for multivariate mcdonald's beta mixture model with feature selection," in *2023 9th International Conference on Control, Decision and Information Technologies (CoDIT)*, 2023, pp. 2055–2060.

[35] J. Wang, Z. Wang, C. Yang, N. Wang, and X. Yu, "Optimization of the number of components in the mixed model using multi-criteria decision-making," *Applied Mathematical Modelling*, vol. 36, no. 9, pp. 4227–4240, 2012.

[36] M. Amirkhani, N. Manouchehri, and N. Bouguila, "Birth-death MCMC approach for multivariate beta mixture models in medical applications," in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices - 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26-29, 2021, Proceedings, Part I*, ser. Lecture Notes in Computer Science, H. Fujita, A. Selamat, J. C. Lin, and M. Ali, Eds., vol. 12798. Springer, 2021, pp. 285–296.

[37] A. Jara, "Theory and computations for the dirichlet process and related models: An overview," *International Journal of Approximate Reasoning*, vol. 81, pp. 128–146, 2017.

[38] N. Bouguila and D. Ziou, "A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling," *IEEE Transactions on Neural Networks*, vol. 21, no. 1, pp. 107–122, 2010.

[39] N. Bouguila, "Infinite liouville mixture models with application to text and texture categorization," *Pattern Recognit. Lett.*, vol. 33, no. 2, pp. 103–110, 2012.

[40] A. Kottas, "Dirichlet process mixtures of beta distributions , with applications to density and intensity estimation," 2006.

[41] M. M. Taye, "Understanding of machine learning with deep learning: Architectures, workflow, applications and future directions," *Computers*, vol. 12, no. 5, 2023.

[42] W. Fan, N. Bouguila, and D. Ziou, "Unsupervised hybrid feature extraction selection for high-dimensional non-gaussian data clustering with variational inference," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1670–1685, 2013.

[43] W. Fan and N. Bouguila, "Unsupervised feature selection for proportional data clustering via expectation propagation," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, 2013, pp. 1–8.

[44] S. Boutemedjet, N. Bouguila, and D. Ziou, "Unsupervised feature and model selection for generalized dirichlet mixture models," in *Image Analysis and Recognition*, M. Kamel and A. Campilho, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 330–341.

[45] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977. [Online]. Available: http://www.jstor.org/stable/2984875

[46] C¸ağlar Arı, S. Aksoy, and O. Arıkan, "Maximum likelihood estimation of gaussian mixture models using stochastic search," *Pattern Recognition*, vol. 45, no. 7, pp. 2804–2816, 2012.

[47] C. Williams, "A mcmc approach to hierarchical mixture modelling," *Advances in Neural Information Processing Systems*, 04 2000.

[48] S. Chib, "Chapter 57 - markov chain monte carlo methods: Computation and inference," ser. Handbook of Econometrics, J. J. Heckman and E. Leamer, Eds. Elsevier, 2001, vol. 5, pp. 3569–3649.

[49] R. Bardenet, A. Doucet, and C. Holmes, "On markov chain monte carlo methods for tall data," *J. Mach. Learn. Res.*, vol. 18, no. 1, p. 1515–1557, Jan. 2017.

[50] W. Fan, N. Bouguila, and D. Ziou, "Variational learning of finite dirichlet mixture models using component splitting," *Neurocomputing*, vol. 129, pp. 3–16, 2014.

[51] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, vol. 112, pp. 859 – 877, 2016.

[52] W. Fan and N. Bouguila, "Learning finite beta-liouville mixture models via variational bayes for proportional data clustering," in *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, F. Rossi, Ed. IJCAI/AAAI, 2013, pp. 1323–1329.

[53] W. Fan, N. Bouguila, and D. Ziou, "Variational learning for finite dirichlet mixture models and applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 5, pp. 762–774, 2012.

[54] W. Fan and N. Bouguila, "Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 11, pp. 1850–1862, 2013.

[55] W. Fan, H. Sallay, and N. Bouguila, "Online learning of hierarchical pitman–yor process mixture of generalized dirichlet distributions with feature selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 9, pp. 2048–2061, 2017.

[56] S. J. Gershman, M. D. Hoffman, and D. M. Blei, "Nonparametric variational inference," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 235–242.

[57] D. Azzam, M. Azam, and N. Bouguila, "Variational inference for the bayesian libby-novick beta mixture model with feature selection," 2025, submitted to [International Journal of Imaging Systems and Technology].

[58] ——, "Nonparametric variational infinite libby-novick beta mixture model for medical data clustering," in *2025 IEEE 34th International Symposium on Industrial Electronics (ISIE)*, 2025, pp. 1–7.

[59] ——, "Nonparametric neural variational inference for mcdonald's beta mixture models with feature selection," 2025, submitted to [Journal of Ambient Intelligence and Humanized Computing].

[60] M. Wand, J. Ormerod, S. Padoan, and R. Fuhrwirth, "Mean field variational bayes for elaborate distributions," *Bayesian Analysis*, vol. 6, 12 2011.

[61] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:174065

[62] C.-F. Tsai, "Bag-of-words representation in image annotation: A review," *International Scholarly Research Notices*, vol. 2012, no. 1, p. 376804, 2012.

[63] N. Tangpukdee, C. Duangdee, P. Wilairatana, and S. Krudsood, "Malaria diagnosis: a brief review," *The Korean journal of parasitology*, vol. 47, no. 2, p. 93, 2009.

[64] W. Siłka, M. Wieczorek, J. Siłka, and M. Wo źniak, "Malaria detection using advanced deep learning architecture," *Sensors*, vol. 23, no. 3, 2023.

[65] Iarunava, "Cell images for detecting malaria," https://www.kaggle.com/datasets/iarunava/cell-images-for-detecting-malaria, 2019.

[66] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 74, no. 3, pp. 229–263, 2024.

[67] J. Chen and J. Dhahbi, "Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods," *Scientific Reports*, vol. 11, p. 13323, 06 2021.

[68] A. A. Borkowski, M. M. Bui, L. B. Thomas, C. P. Wilson, L. A. DeLand, and S. M. Mastorides, "Lung and colon cancer histopathological image dataset (lc25000)," 2019. [Online]. Available: https://arxiv.org/abs/1912.12142

[69] C. E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *The Annals of Statistics*, vol. 2, no. 6, pp. 1152–1174, 1974.

[70] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–, 11 2004.

[71] L. J. Rao, P. Neelakanteswar, M. Ramkumar, A. Krishna, and C. Z. Basha, "An effective bone fracture detection using bag-of-visual-words with the features extracted from sift," in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, pp. 6–10.

[72] D. Blei and M. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, 03 2006.

[73] L. I. Boone, "Chapter 65 - disorders of white blood cells," in *Handbook of Small Animal Practice (Fifth Edition)*, fifth edition ed., R. V. Morgan, Ed. Saint Louis: W.B. Saunders, 2008, pp. 641–655.

[74] R. Gupta, S. Gehlot, and A. Gupta, "C-nmc: B-lineage acute lymphoblastic leukaemia: A blood cancer dataset," *Medical Engineering Physics*, vol. 103, p. 103793, 2022.

[75] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://openreview.net/forum?id=rkE3y85ee

[76] N. Hasan, A. Nadaf, M. Imran, U. Jiba, A. Sheikh, W. Almalki, S. Almujri, M. Hussain, P. Kesharwani, and F. Ahmad, "Skin cancer: understanding the journey of transformation from conventional to advanced treatment approaches," *Molecular Cancer*, vol. 22, 10 2023.

[77] C. Fanconi, "Skin cancer: malignant vs. benign, processed skin cancer pictures of the isic archive," 2019. [Online]. Available: https://www.kaggle.com/fanconic/skin-cancer-malignant-vs-benign

[78] T. Terwilliger and M. Abdul-Hay, "Acute lymphoblastic leukemia: a comprehensive review and 2017 update," *Blood Cancer Journal*, vol. 7, p. e577, 06 2017.

[79] M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, and H. Abolghasemi, "A fast and efficient cnn model for b-all diagnosis and its subtypes classification using peripheral blood smear images," *International Journal of Intelligent Systems*, vol. 37, no. 8, pp. 5113–5133, 2022.