OPEN ACCESS

Artificial Intelligence | PERSPECTIVE

The Innovation Informatics

Informative missingness and its implications in semi-supervised learning

Jinran Wu,[1] You-Gan Wang,[2,3] and Geoffrey J. McLachlan[1,*]

[1]School of Mathematics and Physics, The University of Queensland, St Lucia 4072, Australia

[2]School of Statistics and Data Science, Guangdong University of Finance & Economics, Guangzhou 510320, China

[3]School of Mathematics and Computational Science/Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Xiangtan 411105, China

*Correspondence: g.mclachlan@uq.edu.au

Received: October 28, 2025; Accepted: November 12, 2025; Published Online: January 9, 2026; https://doi.org/10.59717/j.xinn-inform.2026.100033

Citation: Wu J., Wang Y. G. and McLachlan G. J. (2026). Informative missingness and its implications in semi-supervised learning. *The Innovation Informatics* **2**:100033.

Semi-supervised learning (SSL) constructs classifiers using both labelled and unlabelled data. It leverages information from labelled samples, whose acquisition is often costly or labour-intensive, together with unlabelled data to enhance prediction performance. This defines an incomplete-data problem, which statistically can be formulated within the likelihood framework for finite mixture models that can be fitted using the expectation-maximisation (EM) algorithm. Ideally, one would prefer a completely labelled sample, as one would anticipate that a labelled observation provides more information than an unlabelled one. However, when the mechanism governing label absence depends on the observed features or the class labels or both, the missingness indicators themselves contain useful information. In certain situations, the information gained from modelling the missing-label mechanism can even outweigh the loss due to missing labels, yielding a classifier with a smaller expected error than one based on a completely labelled sample analysed.[1] This improvement arises particularly when class overlap is moderate, labelled data are sparse, and the missingness is informative. Modelling such informative missingness thus offers a coherent statistical framework that compares likelihood-based inference with the behaviour of empirical SSL methods.

In many classification tasks, only a subset of class labels is typically observed. In practice, labels are often omitted for ambiguous instances or are recorded preferentially by class. In the terminology of missing-data mechanisms,[2] such missingness is informative. A likelihood-based formulation allows this informative mechanism to be modelled explicitly. The joint distribution of features and class labels can be represented by a finite mixture model, with estimation performed via EM.[3] The pattern of missing labels itself conveys information about where classes overlap and how labels are allocated. Recent analyses have shown that partially labelled data analysed with a correctly specified mechanism can produce more accurate predictions than completely labelled data.[1,4] The focus here is on likelihood structures rather than empirical construction. The relationship between missingness mechanisms and their implications for semi-supervised learning (SSL) is summarised schematically in Figure 1.

LIKELIHOOD FORMULATION FOR PARTIALLY LABELLED DATA

Let $(\mathbf{Y}, Z, M)$ denote the random variables corresponding to the feature vector, class label, and missingness indicator, respectively, where $\mathbf{Y} \in \mathbb{R}^p$, $Z \in \{1,\ldots,g\}$, and $M \in \{0,1\}$ indicates whether the label is missing ($M=1$) or observed ($M=0$). Assume that $\mathbf{Y}$ follows a finite mixture model with g components, $f(\mathbf{y};\boldsymbol{\theta}) = \sum_{i=1}^{g} \pi_i f_i(\mathbf{y};\boldsymbol{\omega}_i)$, where $\pi_i (>0)$ is the class proportion and $\sum_{i=1}^{g} \pi_i = 1$, $f_i(\cdot;\boldsymbol{\omega}_i)$ is the class-conditional density, and $\boldsymbol{\theta} = \{\pi_i, \boldsymbol{\omega}_i\}_{i=1}^{g}$ is the full parameter set. Given data $\{(\mathbf{y}_j, z_j, m_j) : j = 1,\ldots,n\}$, the observed-data log-likelihood can be written as

$$\log L(\boldsymbol{\theta}) = \sum_{j=1}^{n} (1-m_j) \sum_{i=1}^{g} z_{ij} \log\{\pi_i f_i(\mathbf{y}_j;\boldsymbol{\omega}_i)\} + \sum_{j=1}^{n} m_j \log\Big\{\sum_{i=1}^{g} \pi_i f_i(\mathbf{y}_j;\boldsymbol{\omega}_i)\Big\}, \quad (1)$$

where $z_{ij} = 1$ if sample j belongs to class i, and 0 otherwise. The first term corresponds to labelled observations ($M=0$), while the second integrates over unlabelled cases ($M=1$). Where all labels are observed, (1) reduces to the likelihood in the completely labelled situation; where all the labels are missing, it reduces to the standard mixture model-based clustering form. Thus, partially labelled data can be viewed as an intermediate setting between these extremes, where both labelled and unlabelled observations contribute to parameter estimation within a joint likelihood framework. It shall be noted that the posterior class probabilities $\boldsymbol{\tau}_{ij}$ for unlabelled observations $\mathbf{y}_j$ provide fractional estimates of cluster memberships. This likelihood-based formulation is implemented within the EM framework for SSL by which inference with partially labelled data is a natural instance of incomplete-data estimation.[2,5]

INFORMATIVE MISSINGNESS: MECHANISM AND ESTIMATION

In the past, the focus has been on the theoretical results for situations where class missingness is ignorable. In many applications, however, the probability that a label is missing may depend on the data itself and therefore be informative. Two common forms of informative missingness are considered.[2] Under missing at random (MAR), the probability of a label being unobserved depends on the observed feature vector $\mathbf{y}_j$ through a measure of classification uncertainty, such as entropy $e(\mathbf{y}_j;\boldsymbol{\theta}) = -\sum_{i=1}^{g} \tau_{ij} \log \tau_{ij}$. A convenient specification adopts a logistic form, $\operatorname{logit}\{\Pr(M_j = 1 \mid \mathbf{Y} = \mathbf{y}_j)\} = \boldsymbol{\xi}^{\top} \mathbf{u}(\mathbf{y}_j)$, where $\mathbf{u}(\mathbf{y}_j)$ may include $e(\mathbf{y}_j;\boldsymbol{\theta})$ or related features, and $\boldsymbol{\xi}$ denotes the mechanism parameters. This captures feature-dependent missingness: ambiguous instances near class boundaries are less likely to be labelled. Under missing not at random (MNAR), the probability of being unlabelled also depends on the true class label and feature vectors, $\Pr(M_j = 1 \mid Z = i, \mathbf{Y} = \mathbf{y}_j) = \phi_i(b_{\mathbf{y}_j})$, where ϕ_i represent the class-dependent missingness probability evaluted at the discriminant term $b_{\mathbf{y}_j}$ for sample $\mathbf{y}_j$ from class i. This reflects class-dependent or preferential missingness, where annotation effort or accessibility varies systematically across classes. Both MAR and MNAR correspond to non-ignorable mechanisms in the sense of Rubin,[2] and the indicators M_j thereby contain additional information about decision boundaries and class proportions.

From an information-theoretic perspective, the total Fisher information (the negative of the expected curvature of the log-likelihood) in a partially labelled sample can be decomposed into three components.[1] Let $\mathbf{I}_{\mathrm{cc}}$ denote the information that would be available under complete classification, $\mathbf{I}_{\mathrm{uc}}$ the information in the unlabelled data, and let $\mathbf{I}_{\mathrm{cc}}^{(\mathrm{clr})}$ represent the conditional information for logistic regression.[4] The contribution of the missingness mechanism itself is denoted by $\mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})}$, which quantifies the information gained by modelling how labels are missing. The total information can be written schematically as

$$\begin{aligned}\mathbf{I}_{\mathrm{pc}}^{(\mathrm{full})} &= (1-\gamma)\,\mathbf{I}_{\mathrm{cc}} + \gamma\,\mathbf{I}_{\mathrm{uc}} + \mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})} \\ &= (1-\gamma)\,\mathbf{I}_{\mathrm{cc}} + \gamma\,(\mathbf{I}_{\mathrm{cc}} - \mathbf{I}_{\mathrm{cc}}^{(\mathrm{clr})}) + \mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})} \\ &= \mathbf{I}_{\mathrm{cc}} - \gamma\,\mathbf{I}_{\mathrm{cc}}^{(\mathrm{clr})} + \mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})},\end{aligned}$$

where γ is the expected proportion of missing labels. The first term measures the potential information under full supervision; the second subtracts the loss due to incomplete labels; and the third adds the gain from modelling the missingness mechanism. When the mechanism is non-informative, $\mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})} = 0$, and the total information decreases with γ. However, when the missingness depends on features or latent class structure, $\mathbf{I}_{\mathrm{pc}}^{(\mathrm{miss})} > 0$ and can exceed $\gamma\,\mathbf{I}_{\mathrm{cc}}^{(\mathrm{clr})}$, producing the apparent paradox that partially labelled data analysed under a correctly specified mechanism may be more informative than completely labelled data analysed. In such cases, the missing-label indicators effectively serve as auxiliary variables, sharpening inference on the class boundaries and mixing proportions. The magnitude of this gain is greatest where class overlap

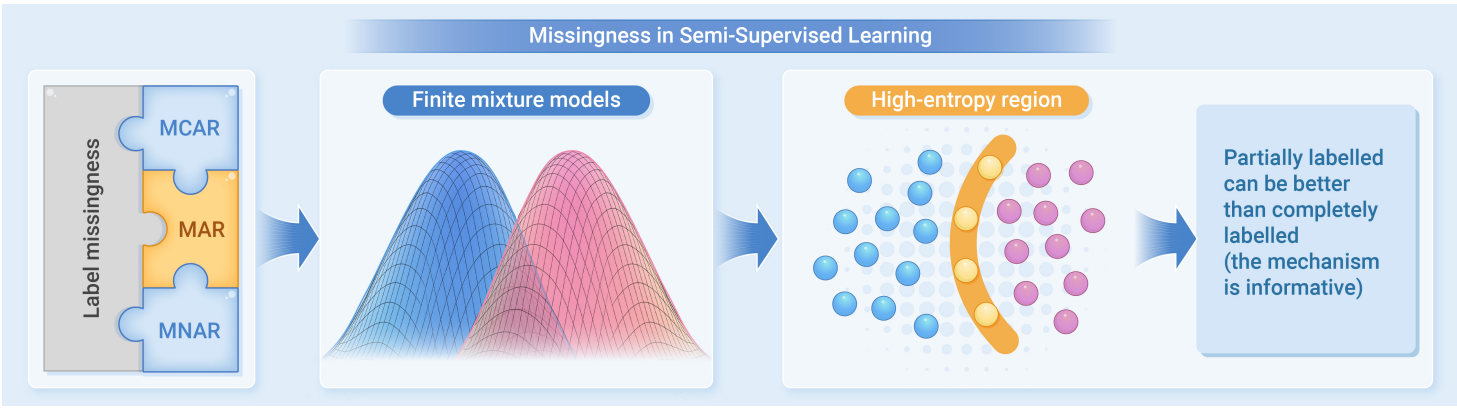

Figure 1. Schematic illustration showing how MCAR, MAR, and MNAR missingness mechanisms influence model-based inference and classification boundaries in SSL When the missingness mechanism is correctly modelled within a finite-mixture framework, a classifier trained on partially labelled data can outperform one trained on completely labelled data.[1]

is moderate, the labelled subset is small, and the missingness mechanism is strongly dependent on feature uncertainty or class membership. Intuitively, this means that when missingness carries useful structure, the absence of some labels can itself provide additional information for learning.

The likelihood approach provides a direct way to formulate models with informative missingness. The observed-data log-likelihood is extended to include the contribution of the missingness indicators M_j, so that both the mixture and mechanism parameters are estimated jointly, typically via the EM algorithm. This allows uncertainty to be propagated coherently and yields efficient estimates when the mechanism is specified. Gains are most visible when class overlap is moderate and labelled data are sparse.

RELATION TO EMPIRICAL SSL

Most SSL algorithms implicitly assume how labels are missing and how uncertainty determines which unlabelled data contribute to learning. Classical SSL theory rests on the smoothness and low-density assumptions, suggesting that decision boundaries should lie in regions of high uncertainty or low sample density. Under this view, minimum-entropy regularisation corresponds to an MAR mechanism, where unlabelled data concentrate near ambiguous regions.[6] This interpretation extends naturally to deep learning methods such as MixMatch,[7] which combines guessed labels and data augmentation to enforce prediction consistency, thereby reducing entropy in uncertain regions. From this standpoint, these approaches can be regarded as empirical realisations of MAR-type mechanisms that regulate the influence of unlabelled data during training-implicitly recognising that the absence of labels is itself informative.

Recent surveys have systematically mapped the evolution of SSL from probabilistic mixture formulations to deep neural frameworks.[8,9] As summarised by Yang et al.,[8] empirical SSL methods fall into five major paradigms: deep generative models, consistency regularisation, graph-based inference, pseudo-labelling, and hybrid frameworks that integrate multiple strategies for state-of-the-art performance. Despite architectural diversity, these paradigms share a common inductive bias: unlabelled data are not treated as MCAR but are weighted by their predictive uncertainty, consistency, or neighbourhood structure conceptually similar to an entropy-driven MAR mechanism. Gui et al.[9] further showed that class imbalance induces systematic bias in pseudo-labelling and consistency-based models, mirroring class-dependent or preferential missingness analogous to an MNAR mechanism. These observations highlight that the performance of deep SSL depends critically on how the implicit label-missingness process is represented. Hence, empirical SSL algorithms can be interpreted as heuristic approximations to informative missingness, where entropy minimisation and distribution alignment serve as empirical surrogates for the statistical mechanisms governing label absence. This perspective provides a conceptual bridge between likelihood-based inference and deep learning practice, offering a foundation for unifying theoretical and empirical SSL within a mechanism-aware framework.

CONCLUDING REMARKS

The central objective is to utilise informative missingness to enhance the predictive performance of classifiers trained on partially labelled data. Missingness arising from regions of high uncertainty within the feature space can provide additional information that enhances model estimation and classification. Within a finite-mixture and EM framework, incorporating a MAR or MNAR mechanism enables the systematic exploitation of this information through a joint-likelihood formulation. The potential benefit is most pronounced in problems with moderate class overlap and limited labelled data. When the mechanism is well specified, the resulting classifier can achieve a smaller expected error than one trained on a completely labelled sample.[1,4] Modelling of the missing-label mechanism is therefore both practical and important for achieving reliable and accurate inference.

Several avenues for future exploration remain open. From a theoretical perspective, further work should aim to establish general information-theoretic bounds that quantify how missingness-dependent uncertainty contributes to Fisher information and generalisation performance, and to formalise the link between entropy-based MAR mechanisms and empirical SSL objectives. Additionally, its sensitivity to model misspecification and robustness within the Godambe information bound merit future investigation. At the methodological level, developing hybrid algorithms that learn or approximate the missingness mechanism, for example, entropy-driven weighting schemes or neural architectures that jointly estimate $\Pr(M \mid \mathbf{Y})$ alongside deep neural network parameters, could further integrate likelihood-based inference with deep SSL. In terms of applications, informative-missingness modelling can be extended to domains such as medical diagnosis, environmental sensing, and geoscientific monitoring, where label incompleteness often reflects systematic or human-driven bias rather than randomness. Embedding informative missingness within the broader SSL paradigm provides a coherent pathway to compare classical statistical modelling and deep learning practice, advancing the theoretical, methodological, and applied frontiers of data-efficient learning.

REFERENCES

1. Ahfock D. and McLachlan G.J. (2020). An apparent paradox: A classifier based on a partially classified sample may have smaller expected error rate than that if the sample were completely classified. *Stat. Comput.* **30**:1779–1790. DOI:10.1007/s11222-020-09971-5
2. Rubin D.B. (1976). Inference and missing data. *Biometrika* **63**:581–592. DOI:10.1093/biomet/63.3.581
3. Dempster A.P., Laird N.M. and Rubin D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**:1–22. DOI:10.1111/j.2517-6161.1977.tb01600.x
4. Ahfock D. and McLachlan G.J. (2023). Semi-supervised learning of classifiers from a statistical perspective: A brief review. *Econ. Stat.* **26**:124–138. DOI:10.1016/j.ecosta.2022.03.007
5. McLachlan G.J., Lee S.X. and Rathnayake S.I. (2019). Finite mixture models. *Annu. Rev. Stat. Appl.* **6**:355–378. DOI:10.1146/annurev-statistics-031017-100325
6. Grandvalet Y. and Bengio Y. (2004). Semi-supervised learning by entropy minimization. *Adv. Neural Inf. Process. Syst.* **17**:529–536. DOI:10.5555/2976040.2976107

7. Sohn K., Berthelot D., Carlini N., et al. (2020). FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Adv. Neural Inf. Process. Syst.* **33**:596–608. DOI:10.5555/3495724.3495775
8. Yang X., Song Z., King I., et al. (2023). A survey on deep semi-supervised learning. *IEEE Trans. Knowl. Data Eng.* **35**:8934–8954. DOI:10.1109/TKDE.2022.3220219
9. Gui Q., Zhou H., Guo N., et al. (2024). A survey of class-imbalanced semi-supervised learning. *Mach. Learn.* **113**:5057–5086. DOI:10.1007/s10994-023-06344-7

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the Australian Research Council [DP230101671], ARC Training Centre on Innovation in Biomedical Imaging Technology [IC170100035], and the Science and Technology Innovation Program of Hunan Province [2022RC4028]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

Jinran Wu drafted the manuscript, and You-Gan Wang and Geoffrey J. McLachlan provided critical revisions and intellectual guidance. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

You-Gan Wang is an Editorial Board member of The Innovation journals and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and their research group. The other authors declare no conflicts of interest.