



*Research Paper***A Generative Cross-Modal Imputation Framework Using Variational Autoencoders for Breast Cancer Prognosis under Incomplete Multimodal Data**¹Sumayya Samreen¹Assistant Professor, Department of Computer Science and Engineering, Methodist College of Engineering and Technology, Abids, Hyderabad, Telangana 500001, India¹Email: sumayyasamreen01@gmail.com**ABSTRACT**

Accurate prognosis of breast cancer supports decisions about adjuvant therapy, surveillance intensity, and patient counselling. Prognostic information is distributed across several data sources, including clinical attributes, messenger RNA (mRNA) expression, and copy number alteration (CNA) profiles, and combining these sources through multimodal deep learning can improve predictive performance. In practice, however, multimodal datasets are rarely complete. Genomic profiling may not be performed for every patient, samples may fail quality control, and records collected across institutions frequently differ in the data they contain. Conventional solutions either discard incomplete patients or apply simple imputation methods that ignore the relationships between modalities. This paper proposes "A Generative Cross-Modal Imputation Framework Using Variational Autoencoders for Breast Cancer Prognosis under Incomplete Multimodal Data". The framework trains a separate variational encoder for each modality and combines the observed modalities in a shared latent space using a product-of-experts formulation. Because the joint latent representation can be inferred from any subset of observed modalities, it can be decoded to reconstruct the missing modalities, providing biologically informed imputation instead of placeholder values. The completed latent representation is then used by a prognosis classifier to predict five-year survival status and recurrence risk, while Monte Carlo dropout is used to estimate prediction uncertainty, which is expected to increase when more modalities are missing. The framework can be evaluated using accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve, calibration error, and imputation reconstruction error under controlled missing-data rates. The proposed approach aims to make full use of incomplete breast cancer cohorts, although its outputs require independent clinical validation before use in medical practice.

Keywords: Breast cancer prognosis, missing data, variational autoencoder, cross-modal imputation, product of experts, multimodal deep learning, gene expression, uncertainty estimation.

I. INTRODUCTION

Breast cancer represents a major global health burden and is the most commonly diagnosed cancer among women in most countries [1]. Treatment decisions depend strongly on the expected course of the disease, and therefore reliable prognostic models are needed to identify patients who may benefit from more aggressive therapy and those who may safely avoid unnecessary treatment. The prognosis of breast cancer is

influenced by clinical characteristics, such as tumour size, grade, nodal status, and receptor status, as well as by molecular events that are captured in genomic and transcriptomic data. Lobato-Delgado et al. [2] reviewed the combination of molecular, imaging, and clinical data for predicting cancer prognosis and concluded that integrated analysis provides a more complete picture of tumour behaviour than any single source.

Multimodal deep learning has been widely applied to breast cancer prognosis. Sun et al. [3] proposed a multimodal deep neural network that integrated gene expression, copy number alteration, and clinical data to predict long-term and short-term survivors among breast cancer patients. Arya and Saha [4] proposed a deep learning-based stacked ensemble model that combined clinical, gene expression, and copy number data for breast cancer prognosis prediction. These studies show that each modality contributes complementary information and that their combination improves prediction. However, most such models are trained and evaluated only on patients with all modalities available.

In practice, missing modalities are common. A patient may have complete clinical records but no gene expression profile, or a genomic profile without corresponding copy number data. Removing such patients reduces the amount of training data and can introduce selection bias. Classical imputation techniques have been developed for missing values in tabular data. Van Buuren and Groothuis-Oudshoorn [5] described multivariate imputation by chained equations, which iteratively models each variable from the others. Yoon et al. [6] proposed GAIN, a generative adversarial network for missing data imputation. These methods are effective for scattered missing values, but they are less suited to block-wise missingness, in which an entire high-dimensional modality is absent for a patient.

Deep generative models offer a principled way to learn the joint distribution of multiple modalities. Kingma and Welling [7] introduced the variational autoencoder (VAE), which learns a probabilistic latent representation of data and can generate new samples from that representation. Wu and Goodman [8] extended this idea to multiple modalities using a product-of-experts inference network, so that a joint latent representation can be inferred from any subset of observed modalities. Tran et al. [9] proposed a cascaded residual autoencoder for imputing missing modalities, demonstrating that the correlation between modalities can be exploited to reconstruct absent data. Ma et al. [10] studied multimodal learning when a modality is missing for most training samples

and used Bayesian meta-learning to approximate the missing modality.

Missing-modality methods have also been explored in cancer prognosis. Cui et al. [11] investigated survival prediction of brain cancer using incomplete radiology, pathology, genomic, and demographic data, and showed that reconstructing missing modalities in a latent space can improve predictions for incomplete patients. Huang et al. [12] developed SALMON, a survival analysis model that integrated multi-omics data for breast cancer using co-expression network modules, while Liu et al. [13] proposed a hybrid deep learning model that used multimodal data to predict molecular subtypes of breast cancer. These studies demonstrate the value of multimodal molecular data in breast cancer but also highlight the need for robust methods when some data are unavailable.

An additional consideration is the reliability of predictions made from incomplete data. A prediction based only on clinical attributes should generally be less certain than a prediction based on clinical, transcriptomic, and copy number data. Gal and Ghahramani [14] showed that dropout in neural networks can be interpreted as approximate Bayesian inference, allowing prediction uncertainty to be estimated by performing multiple stochastic forward passes. Combining imputation with uncertainty estimation can therefore help indicate how much confidence should be placed in each prediction.

To address these challenges, this paper proposes a Generative Cross-Modal Imputation Framework that uses multimodal variational autoencoders for breast cancer prognosis under incomplete data. Each modality is encoded into a Gaussian distribution in a shared latent space, and the distributions of the observed modalities are combined using a product of experts. The joint latent representation is used both to reconstruct missing modalities and to predict prognosis. The framework can be applied to cohorts such as METABRIC [15], which contains clinical, gene expression, and copy number data for breast cancer patients, by simulating different missing-modality patterns.

The primary objective of this research is to investigate whether generative cross-modal imputation in a shared latent space can maintain prognostic performance when one or more modalities are missing, while providing uncertainty estimates that reflect the amount of available information. The outputs are intended to support research and computer-aided prognosis and are not a substitute for clinical diagnosis.

II. LITERATURE SURVEY

Van Buuren and Groothuis-Oudshoorn (2011) described Multivariate Imputation by Chained Equations (MICE), an approach that fills missing values by iteratively modelling each incomplete variable as a function of the other variables. MICE is flexible and widely used in medical research for handling scattered missing values in tabular data. However, when an entire high-dimensional modality, such as a gene expression profile with thousands of genes, is missing for a patient, chained regression models become computationally demanding and may fail to capture the complex non-linear relationships between modalities.

Yoon et al. (2018) proposed GAIN, a generative adversarial imputation network in which a generator imputes missing components and a discriminator attempts to identify which components were imputed. A hint mechanism guides the discriminator so that the generator learns the true data distribution. GAIN demonstrated improved imputation performance compared with conventional methods. However, it was designed mainly for missing entries within a single feature vector, and adversarial training can be unstable when the data are high-dimensional and the sample size is limited.

Wu and Goodman (2018) introduced the multimodal variational autoencoder (MVAE), which uses a product-of-experts inference network to combine the latent distributions of individual modalities. This formulation allows the model to infer a joint latent representation from any subset of modalities and to generate missing modalities from that representation. The approach scales efficiently to many modalities and supports weakly supervised learning. Its effectiveness has been

demonstrated mainly on image and text benchmarks, and its application to clinical prognosis remains an open research direction.

Tran et al. (2017) proposed a cascaded residual autoencoder for missing modality imputation, in which a series of residual autoencoders progressively refine the reconstruction of missing data. The method exploits correlations between modalities to estimate the absent information and showed that imputed modalities can support downstream classification. However, the deterministic autoencoder does not model uncertainty in the imputed values, which is important when the imputation is used for medical prediction.

Arya and Saha (2022) proposed a multimodal classification approach for human breast cancer prognosis prediction based on a deep learning-based stacked ensemble model. Convolutional neural networks extracted features from clinical, gene expression, and copy number data, and the resulting features were combined using a stacked ensemble classifier. The approach showed improved prognosis prediction compared with single-modality models. However, the model assumes complete data for all patients and does not address situations in which one or more modalities are missing.

Cui et al. (2022) investigated survival prediction for brain cancer with incomplete radiology, pathology, genomic, and demographic data. Their approach used a modality-specific latent representation and reconstruction strategy so that patients with missing modalities could still be included in training and prediction. The study demonstrated that latent-space reconstruction of missing modalities is a promising strategy for cancer prognosis. However, it focused on brain cancer, and its applicability to breast cancer cohorts and molecular data types requires further investigation.

III. EXISTING SYSTEM

The existing systems for breast cancer prognosis mainly use statistical models and machine learning classifiers trained on

clinical attributes, sometimes combined with gene expression or copy number data. Multimodal deep learning systems typically use separate feature extractors for each modality and combine their outputs through feature concatenation or ensemble learning to predict outcomes such as five-year survival or recurrence. These approaches have shown that combining clinical and molecular data improves prognostic performance compared with clinical data alone.

However, existing multimodal systems usually require complete data for every patient. In practice, patients with missing modalities are frequently excluded from the analysis, which reduces the size of the training set and may bias the cohort. When imputation is used, it is usually performed with simple methods such as mean replacement, zero filling, or nearest-neighbour imputation applied independently to each feature. These methods ignore the relationships between modalities, are poorly suited to entire missing modalities, and can produce unrealistic values that distort the downstream model. In addition, most existing systems provide a single point prediction without indicating how confident the model is, even when the prediction is based on much less information than usual. Therefore, existing systems do not adequately address block-wise missing modalities or provide reliable estimates of prediction confidence.

IV. PROPOSED SYSTEM

The proposed system introduces a Generative Cross-Modal Imputation Framework using variational autoencoders for breast cancer prognosis under incomplete multimodal data. The framework considers three modalities: clinical attributes, mRNA expression, and copy number alteration profiles. Each modality is processed by a dedicated variational encoder that maps it to the mean and variance of a Gaussian distribution in a shared latent space. For each patient, the distributions of the observed modalities are combined with a standard Gaussian prior using a product of experts, producing a joint latent distribution. Because the product of experts can be computed from

any subset of modalities, the framework does not require all modalities to be present.

The joint latent representation is decoded by modality-specific decoders to reconstruct both observed and missing modalities. The reconstruction of observed modalities trains the generative model, while the reconstruction of missing modalities provides cross-modal imputation that reflects learned relationships between clinical, transcriptomic, and copy number data. The joint latent representation, together with a summary of the imputed features, is passed to a prognosis classifier that predicts five-year survival status and recurrence risk. The generative and predictive components are trained jointly so that the latent space captures information that is useful for prognosis. Monte Carlo dropout is applied at inference to estimate predictive uncertainty. The framework can be evaluated using accuracy, precision, recall, F1-score, AUC-ROC, expected calibration error, and imputation error under different missing-modality rates. The predictions are intended to support research and computer-aided prognosis rather than replace professional medical judgement.

V. SYSTEM ARCHITECTURE

The system architecture of the proposed Generative Cross-Modal Imputation Framework consists of several interconnected modules designed to learn from incomplete multimodal breast cancer data and produce prognosis predictions with associated uncertainty.

The first module is the Incomplete Input module, which contains clinical data, mRNA expression profiles, and copy number alteration profiles, together with prognosis labels such as five-year survival status and recurrence status. For each patient, one or more modalities may be missing. Cohorts such as METABRIC can be used, and controlled missingness can be introduced by removing modalities from selected patients to evaluate robustness.

The second module is Data Preparation. Clinical attributes are cleaned, categorical variables are encoded, and numerical variables are scaled. Gene expression and

copy number features are filtered to retain the most variable genes, and the values are standardised using statistics computed from the training data only. An observed-modality flag is created for each patient to indicate which modalities are available.

The third module is the Modality VAE Encoders module. A separate encoder network is used for each modality, and each encoder outputs the mean and log-variance of a Gaussian distribution over a shared latent space. Missing modalities do not pass through their encoders, which prevents artificial values from influencing the latent representation.

The fourth module is the Product of Experts module. The Gaussian distributions produced by the encoders of the observed modalities are combined with a standard normal prior expert. The product of Gaussian experts is itself Gaussian, with a precision equal to the sum of the individual precisions, so that modalities with more confident encodings contribute more strongly. A latent vector is sampled using the reparameterisation trick.

The fifth module is Cross-Modal Imputation. Modality-specific decoders reconstruct each modality from the joint latent vector. For observed modalities, reconstruction loss is calculated against the true data. For missing modalities, the decoded output serves as the imputed modality. Training with random modality masking encourages the latent space to support reconstruction of any modality from any subset of the others.

The sixth module is the Prognosis Classifier. The joint latent vector is passed through fully connected layers with dropout to predict five-year survival status and recurrence risk using sigmoid outputs. At inference, multiple stochastic forward passes are performed with dropout enabled, and the mean of the outputs is used as the prediction while the variance indicates uncertainty.

Finally, the Evaluation module measures classification performance using accuracy, precision, recall, F1-score, and AUC-ROC, imputation quality using mean squared error between imputed and held-out true modalities, and reliability using expected

calibration error and uncertainty analysis. Performance is compared across missing-modality scenarios and against complete-case, mean-imputation, and MICE-based baselines.

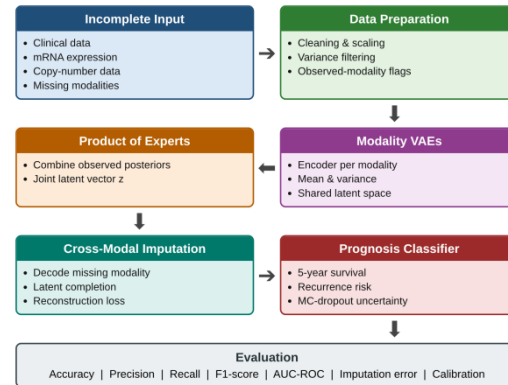


Fig 5.1 System Architecture

VI. ALGORITHM

Algorithm: Generative Cross-Modal Imputation with Product-of-Experts VAE for Breast Cancer Prognosis (GCMi-BCP)

Input:

$D \leftarrow$ Breast cancer dataset with clinical data x_1 , mRNA expression x_2 , copy number data x_3 , and prognosis labels y .

$O_i \leftarrow$ Set of observed modalities for patient i .

$\beta \leftarrow$ Weight of the Kullback–Leibler divergence term.

$\lambda \leftarrow$ Weight of the prognosis loss.

$S \leftarrow$ Number of Monte Carlo dropout samples.

$E_{\max} \leftarrow$ Maximum number of training epochs.

Output:

$M^* \leftarrow$ Trained generative imputation and prognosis model.

$\hat{x}_k \leftarrow$ Imputed values for missing modalities.

$\hat{y}, u \leftarrow$ Predicted prognosis probability and its uncertainty.

Steps:

1. Start.
2. Load the dataset D and determine the set of observed modalities O_i for each patient.

3. Divide the patients into training, validation, and testing subsets using a stratified split based on the prognosis label.

4. Clean and encode clinical attributes, filter and standardise genomic features, and compute all preprocessing statistics from the training subset only.

5. Initialise the encoder q_k and decoder p_k for each modality k , and the prognosis classifier h .

6. For each mini-batch during training, randomly hide a subset of the observed modalities to simulate additional missingness, while keeping at least one modality observed.

7. For each observed modality k , compute the mean μ_k and variance σ_k^2 of its latent distribution using encoder q_k .

8. Combine the observed modality distributions with the prior $N(0, I)$ using the product of experts:

$$\sigma^{-2} = I + \sum_{k \in O} \sigma_k^{-2}, \quad \mu = \sigma^2 \sum_{k \in O} \mu_k \sigma_k^{-2}$$

9. Sample the joint latent vector using the reparameterisation trick:

$$z = \mu + \sigma \odot \varepsilon, \quad \varepsilon \sim N(0, I)$$

10. Decode z using every decoder p_k to obtain reconstructions $\hat{x}_k$ for all modalities.

11. Calculate the reconstruction loss for the observed modalities, including the modalities hidden in step 6, and the Kullback–Leibler divergence between the joint posterior and the prior.

12. Predict the prognosis probability $\hat{y} = h(z)$ and calculate the binary cross-entropy loss with the true label y .

13. Combine the losses into the total training objective:

$$L = \sum_k \|x_k - \hat{x}_k\|^2 + \beta \cdot KL(q(z|x_O) \parallel N(0, I)) + \lambda \cdot BCE(y, \hat{y})$$

14. Update all model parameters using the Adam optimiser.

15. Evaluate AUC-ROC on the validation subset after each epoch and apply early stopping when it stops improving.

16. Save the best model M^* .

17. For each test patient, compute the joint latent distribution from the available modalities, decode the missing modalities to obtain $\hat{x}_k$, and perform S stochastic forward passes through h with dropout enabled.

18. Compute the predicted probability as the mean and the uncertainty as the variance of the S outputs:

$$\hat{y} = (1/S) \sum_s \hat{y}_s, \quad u = (1/S) \sum_s (\hat{y}_s - \hat{y})^2$$

19. Calculate accuracy, precision, recall, F1-score, AUC-ROC, calibration error, and imputation error for each missing-modality scenario.

20. Return the imputed modalities, predicted prognosis, uncertainty, and evaluation results.

21. Stop.

Computational Complexity

Let N be the number of training patients, E the number of epochs, K the number of modalities, and C the average cost of one encoder and decoder pass. Because the product of experts requires only element-wise operations over the latent dimension d , the training complexity is approximately

$$O(NE(KC + Kd))$$

Inference requires one encoding of the observed modalities and S passes through the lightweight classifier, giving $O(KC + SH)$, where H is the classifier cost. Since S and K are small, the framework remains efficient for typical cohort sizes.

VII. CONCLUSION

The proposed Generative Cross-Modal Imputation Framework addresses the problem of block-wise missing modalities in breast cancer prognosis by learning a shared latent space in which clinical, transcriptomic, and copy number data are jointly represented. A product-of-experts formulation enables the joint representation to be inferred from any available subset of modalities, and modality-specific decoders reconstruct missing data using learned cross-modal relationships rather than placeholder values. The prognosis classifier uses the joint representation to predict five-year survival and recurrence risk, while Monte Carlo dropout provides uncertainty estimates that

can indicate when predictions are based on limited information. By using incomplete patients instead of excluding them, the framework can make better use of available data and reduce selection bias. Its effectiveness can be assessed using classification, calibration, and imputation metrics across different missing-data scenarios and in comparison with conventional imputation baselines. Overall, the framework offers a practical research direction for robust multimodal breast cancer prognosis, although rigorous experimental evaluation and independent clinical validation are required before any clinical use.

VIII. FUTURE SCOPE

The future scope of the proposed framework includes extending the generative model to additional modalities, such as histopathology images, mammography, DNA methylation, and proteomic data, which would require modality-appropriate encoders and decoders. Time-to-event survival objectives can be incorporated to model censored follow-up data directly instead of fixed-horizon classification. More expressive generative models, such as hierarchical VAEs or diffusion-based imputation, could improve the quality of reconstructed modalities. Interpretation methods could be applied to the latent space and decoders to identify which genes and clinical variables drive prognosis and imputation. Validation on independent, multi-institutional cohorts with naturally occurring missingness, rather than only simulated missingness, will be important for assessing generalisability. Finally, fairness evaluation across patient subgroups and prospective clinical studies will be necessary before the framework can be considered for practical decision support.

REFERENCES

- [1] F. Bray et al., "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. doi: 10.3322/caac.21834.
- [2] B. Lobato-Delgado, B. Priego-Torres, and D. Sanchez-Morillo, "Combining molecular, imaging, and clinical data analysis for predicting cancer prognosis," *Cancers*, vol. 14, no. 13, Art. no. 3215, 2022. doi: 10.3390/cancers14133215.
- [3] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 16, no. 3, pp. 841–850, 2019. doi: 10.1109/TCBB.2018.2806438.
- [4] N. Arya and S. Saha, "Multi-modal classification for human breast cancer prognosis prediction: Proposal of deep-learning based stacked ensemble model," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 19, no. 2, pp. 1032–1041, 2022. doi: 10.1109/TCBB.2020.3018467.
- [5] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. doi: 10.18637/jss.v045.i03.
- [6] J. Yoon, J. Jordon, and M. van der Schaar, "GAIN: Missing data imputation using generative adversarial nets," in *Proc. 35th International Conference on Machine Learning (ICML)*, PMLR vol. 80, pp. 5689–5698, 2018.
- [7] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. International Conference on Learning Representations (ICLR)*, 2014.
- [8] M. Wu and N. Goodman, "Multimodal generative models for scalable weakly-supervised learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, 2018.
- [9] L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing modalities imputation via cascaded residual autoencoder," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1405–1414, 2017.
- [10] M. Ma et al., "SMIL: Multimodal learning with severely missing modality," in *Proc. AAAI Conference on Artificial*

Intelligence, vol. 35, no. 3, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

[11] C. Cui et al., "Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, pp. 626–635, 2022.

[12] Z. Huang et al., "SALMON: Survival analysis learning with multi-omics neural networks on breast cancer," Frontiers in Genetics, vol. 10, Art. no. 166, 2019. doi: 10.3389/fgene.2019.00166.

[13] T. Liu, J. Huang, T. Liao, R. Pu, S. Liu, and Y. Peng, "A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data," IRBM, vol. 43, no. 1, pp. 62–74, 2022. doi: 10.1016/j.irbm.2020.12.002.

[14] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. 33rd International Conference on Machine Learning (ICML), PMLR vol. 48, pp. 1050–1059, 2016.

[15] C. Curtis et al., "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups," Nature, vol. 486, pp. 346–352, 2012. doi: 10.1038/nature10983.