A Class-Weighted Deep Learning Framework for Multimodal Breast Cancer Prognosis Prediction

Dr. K. KrishnaKumari [*], Ram Prakash S [**], Siddarth G [**], Gowtham K [**], Dinesh M [**]

*Assistant Professor -Computer Science and Engineering, University College of Engineering, Panruti

[*] UG Students -Computer Science and Engineering, University College of Engineering, Panruti

ABSTRACT

Breast cancer prognosis prediction plays an important role in guiding treatment strategies and improving patient survival outcomes. Although several machine learning models have been proposed, many existing approaches struggle with high-dimensional genomic data, limited sample sizes, and imbalanced survival classes. To address these challenges, this study proposes a reproducible two-stage multimodal learning framework for five-year breast cancer survival prediction using clinical variables, gene expression profiles, and copy number alteration (CNA) data derived from the METABRIC cohort. In the first stage, compact one-dimensional convolutional neural networks (CNNs) are trained independently for each modality to learn informative latent feature representations. In the second stage, the learned hidden representations are combined into a unified feature space and provided to ensemble classifiers including Support Vector Machine (SVM), Random Forest, XGBoost, and CatBoost. Instead of applying aggressive data resampling, class imbalance is handled through classifier-level weighting to preserve the original distribution of the dataset. To control the dimensionality of genomic inputs, a two-step feature selection strategy is applied using minimum-redundancy maximum-relevance (mRMR) followed by incremental AUC-guided refinement. This process reduces the input space to 400 gene expression features and 200 CNA features while retaining key clinical attributes. Model evaluation is conducted using stratified ten-fold cross-validation with detailed analysis of discrimination, calibration, and statistical robustness. Experimental results show that the proposed stacked architecture significantly improves prediction performance compared with unimodal and conventional fusion models, achieving a maximum accuracy of 0.93 and ROC-AUC values approaching 0.96. These findings demonstrate that combining modality-specific deep representations with ensemble classifiers provides an effective and reproducible solution for multimodal breast cancer prognosis modelling.

Index Terms—Breast cancer prognosis, Multi-modal learning, Convolutional neural networks, Stacked ensemble, Feature selection, mRMR, Gene expression, Copy-number alteration, Class imbalance, Calibration, Gradient boosting, Model interpretability

1 INTRODUCTION

Accurate prediction of breast cancer survival remains a major objective in precision oncology. Although clinical staging systems provide valuable information regarding disease severity, patients with similar clinical characteristics may experience substantially different outcomes. As a result, there is increasing interest in computational methods that integrate diverse biological and clinical information to improve prognostic modelling.

Large-scale biomedical datasets have enabled the development of data-driven approaches for survival prediction. In particular, resources such as the METABRIC cohort provide detailed clinical annotations together with genomic measurements including gene expression and copy number alterations. These multimodal datasets offer an opportunity to explore complex biological relationships that cannot be captured by traditional clinical indicators alone.

However, modelling such heterogeneous data introduces several challenges. Genomic measurements typically contain thousands of variables while the number of available patient samples remains relatively limited. This imbalance between feature dimensionality and sample size increases the risk of overfitting and reduces model generalization. In addition, prognosis datasets frequently exhibit class imbalance, where long-term survivors significantly outnumber short-term survivors. Conventional learning algorithms may therefore become biased toward the majority class.

Recent advances in deep learning have shown potential for extracting meaningful patterns from high-dimensional biomedical signals. Convolutional neural networks (CNNs), when applied to structured genomic data, can identify local feature relationships and produce compact representations suitable for downstream learning tasks. At the same time, classical machine learning algorithms such as support vector machines(SVM) and tree-based ensembles remain highly effective for structured tabular data and often provide robust decision boundaries.

Motivated by these observations, this work introduces a two-stage multimodal prognosis framework that combines CNN-based representation learning with ensemble classification. The proposed pipeline first learns modality-specific feature representations from clinical, gene expression, and CNA data. These learned features are then integrated and provided to multiple ensemble classifiers that perform the final survival prediction. By combining deep feature extraction with robust ensemble learning, the proposed approach aims to improve prediction accuracy while maintaining model stability and reproducibility.

2 LITERATURE REVIEW

This reviews key prior work on breast-cancer prognosis using uni-modal and multi-modal data, feature-selection strategies, deep representation learning (especially CNNs), stacked and ensemble classifiers, and imbalance-aware methods—highlighting gaps our two-stage CNN to stacked-ensemble pipeline addresses.

Breast cancer prognosis has long benefited from molecular signatures and clinical predictors. Early landmark studies like N. Arya et al.. demonstrated that gene-expression panels can stratify patient survival and outperform some clinical models, establishing the feasibility of molecular prognostics [1],[2]. Subsequent work validated and extended these findings across cohorts, emphasizing the need for robust feature selection and cross-study validation to avoid overfitting in high-dimensional genomic spaces [3],[4].

Feature selection and dimensionality reduction have been central to genomic prognostics. Mutual-information-based methods such as mRMR (minimum-redundancy, maximum-relevance) became popular for selecting compact, informative gene sets from tens of thousands of probes, improving downstream classifier stability [5]. Wrapper and embedded methods (recursive feature elimination, tree-based importance) were later combined with mRMR to refine signatures for prognosis tasks [6].

The rise of deep learning introduced representation learning as an alternative to hand-crafted feature selection. S. U. Kandan et al.. used an One-dimensional convolutional neural networks and compact feedforward DNNs have been applied to ordered genomic and CNA vectors, showing improved local pattern discovery and hierarchical feature extraction compared with shallow models [7],[8]. Studies comparing CNNs and DNNs report that CNNs often produce more expressive hidden representations for structured biomedical signals, particularly when sample sizes are modest and architectures are kept compact to avoid overfitting [9],[10].

Multi-modal integration—combining clinical, gene-expression, and CNA data—has been shown to improve prognostic accuracy versus uni-modal models by leveraging complementary information across modalities [11],[12]. Approaches vary from early fusion (concatenating raw features) to score-level fusion and hybrid deep architectures. Notably, MDNNMD and related multimodal deep frameworks demonstrated gains by learning modality-specific representations and fusing them, but some rely on manually tuned fusion weights or simple DNN extractors, which can limit generalizability [13],[14].

Ensemble and stacked classifiers remain strong baselines for tabular biomedical data. Random forests, SVMs, and gradient-boosted trees (XGBoost, CatBoost) consistently perform well when supplied with informative features stacking hidden representations from deep extractors into classical learners often yields the best of both worlds rich learned features plus robust decision boundaries [15]-[17]. Recent work shows that gradient boosting frameworks, with careful handling of categorical features and sampling (e.g., CatBoost) can match or exceed deep models on structured tasks [18],[19].

A persistent challenge is class imbalance in survival datasets: minority short-term survivors are clinically critical but underrepresented. Techniques include resampling, synthetic oversampling, and cost-sensitive learning; class-weighted SVMs and ensemble methods have proven effective at improving minority recall without distorting data distributions [20].

Gaps and motivation. Prior studies either

- Fuse raw modalities without modality-aware feature extraction.
- Use dnns without exploiting CNN's' local pattern learning for genomic signals.
- Apply stacking without explicit imbalance handling. Our work addresses these gaps by using modality-specific CNNs to extract hidden features, concatenating them into a stacked representation, and evaluating a broad set of class-weighted and boosted classifiers under stratified cross-validation to quantify gains and robustness.

3 METHODOLOGY

3.1 DATASET

Experiments use the pre-processed METABRIC cohort with 1,980 patients after quality control and exclusion of samples with incomplete survival labels. Patients are dichotomized using a five-year cutoff: long-term survivors (>5 years) and short-term survivors (≤5 years). The cohort is imbalanced with 1,489 long-term and 491 short-term survivors. A small subset of patients censored within five years were conservatively assigned to the long-term class following prior practice to avoid discarding limited censored samples.

Each patient is represented by three modalities. Clinical covariates include demographic and pathological variables such as age at diagnosis, tumour size, lymph-node

status, grade, inferred menopausal state and treatment indicators, the clinical vector contains 25 features after preprocessing. Gene-expression measurements are high dimensional, following mRMR filtering and incremental AUC-guided refinement, the gene-expression modality is reduced to 400 features. CNA features are discretized into standard categories and reduced to 200 features. These final counts balance representational richness with statistical tractability given the cohort size.

3.2 PREPROCESSING AND IMPUTATION

Missing genomic values were imputed using a weighted nearest-neighbour approach that preserves local structure in the high-dimensional space. Clinical covariates with missing entries were imputed using median for continuous variables and mode for categorical variables, followed by one-hot encoding where appropriate. Clinical features were min–max scaled to [0,1]. Gene-expression values were standardized or discretized consistent with prior METABRIC preprocessing when required by downstream models. CNA features retained discrete coding.

3.3 FEATURE SELECTION

High dimensionality and limited sample size motivate a conservative two-stage feature selection pipeline. Stage 1 applies mRMR independently to gene and CNA modalities to rank features by relevance and redundancy, producing candidate lists. Stage 2 performs incremental AUC-guided refinement: starting with the top 100 features, a lightweight classifier is trained on training folds and evaluated on validation folds, the feature set is expanded in increments, and the AUC curve is monitored. The final selection chooses the feature count at or near the plateau where additional features yield diminishing returns. For METABRIC this procedure produced stable optima at 400 gene features and 200 CNA features shown in table 1. Selection stability was assessed across multiple random seeds and folds, features with high selection frequency were prioritized for interpretability analyses.

TABLE 1
The Overall Information of Metabric Breast Cancer

Dataset Cut-off (years)	5
Total no. of patients	1980
Long-time survivors	1489
Short-time survivors	491
Median age in diagnosis	61
Average survival (months)	125.1

3.4 STAGE ONE CNNS

For each modality, a compact one-dimensional CNN was trained to extract modality-specific hidden representations. The architecture was intentionally kept small to mitigate overfitting: a single convolutional layer with four filters, kernel size of 15, stride of 2, and same padding, followed by flattening. This was succeeded by a dense hidden layer with 150 units and tanh activation, dropout (0.5 for gene and CNA pipelines), and a sigmoid output for binary classification. Glorot normal initialization was used, with bias initialized to 0.1 and L2 regularization applied to dense weights. Training employed the Adam optimizer (learning rate = $1x10^{-3}$), binary cross-entropy loss, batch size of 8, and up to 20 epochs with early stopping monitored on validation AUC. The 150-dimensional penultimate layer activations were extracted as hidden features for downstream fusion and accuracy values are detailed in table 2.

TABLE 2
Performance of CNNs across modalities under stratified ten-fold cross-validation.

CNN	ACCURACY
Clinical CNN	0.7823 ± 0.0124
CNA CNN	0.7601 ± 0.0142
Expr CNN	0.8015 ± 0.0219

3.5 STAGE TWO STACKED ENSEMBLE

Hidden vectors from the three modality CNNs are concatenated to form a stacked feature vector for each sample and accuracy metrics reaches an targeted accuracy shown in table 3. Stage-Two classifiers evaluated include support vector machines with RBF kernel and class weighting, random forests, with class balancing, and gradient boosting frameworks XGBoost, and CatBoost. Hyperparameters are tuned on inner validation folds with modest grids to avoid overfitting. Classifier-level class weighting or scale_pos_weight is preferred over resampling to preserve the original data distribution while penalizing minority-class errors.

TABLE 3
Comparative performance metrics of stacked ensemble classifiers

Model	Accuracy	Precision	Recall	AUC
CatBoost	**0.931**	0.949	0.764	0.960
XGBoost	0.917	0.873	0.780	0.959
SVM	0.909	0.791	0.862	0.961
Random Forest	0.901	0.894	0.683	0.938

3.6 EVALUATION PROTOCOL

We use stratified ten-fold cross-validation as the primary evaluation protocol. Within each outer training set we perform an 80/20 stratified split for validation and hyperparameter tuning. All selection and tuning steps are

performed strictly inside training folds to avoid leakage. Metrics reported per fold include Accuracy, Precision, Recall (Sensitivity), Specificity, F1, Matthew's Correlation Coefficient and ROC-AUC. Aggregated results are reported as mean ± standard deviation across folds. Paired statistical tests on fold-wise AUCs assess significance, multiple comparisons are controlled where applicable. Calibration is evaluated using calibration curves and Brier scores, Platt scaling or isotonic regression is applied on validation folds when necessary. Operating-point analyses illustrate sensitivity/specificity trade-offs at clinically relevant thresholds.

3.7 EXPERIMENTAL DESIGN

All experiments are implemented in Keras/TensorFlow for CNNs and scikit-learn, XGBoost, and CatBoost for Stage-Two learners. Random seeds are fixed for NumPy, TensorFlow and OS-level randomness to ensure reproducibility. Per-fold predictions, selected feature indices and model checkpoints are saved. Hardware includes GPU acceleration for CNN training and multi-core CPU for ensemble training. Hyperparameter searches are constrained and performed only within training folds.

Ablation studies evaluate the contribution of each design choice. We compare hidden-feature stacking versus output-level stacking, CNN extractors versus shallow DNNs of comparable parameter counts and classifier-level class weighting versus resampling strategies. Sensitivity analyses vary gene and CNA feature counts around the chosen optima and perturb key hyperparameters to assess stability. Robustness checks include repeating the full pipeline with multiple random seeds and comparing alternative imputation methods. The model architecture was presented in figure 1.

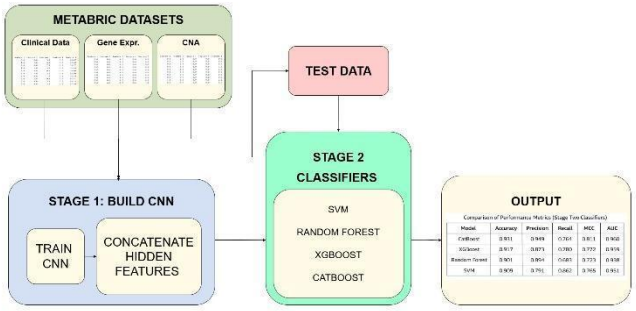

Fig 1. Stacked-based ensemble model architecture

4 RESULTS AND DISCUSSION

4.1 STAGE ONE MODALITY PERFORMANCE

Under stratified ten-fold cross-validation the modality CNNs achieved the following mean accuracies are presented in figure 2.

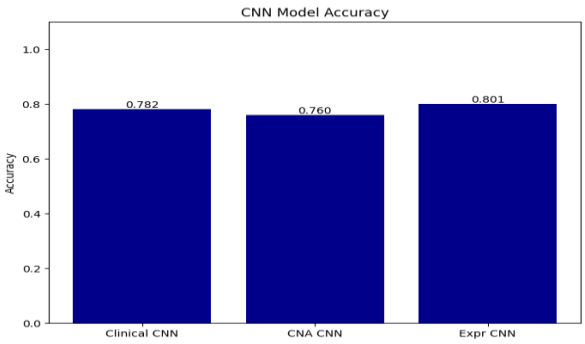

Fig 2. Performance of CNNs

Gene expression provided the strongest single-modality signal, consistent with prior literature.

4.2 STAGE TWO STACKED CLASSIFIERS

Concatenated hidden features fed to Stage-Two classifiers produced substantial gains. Representative aggregated results across folds are summarized below in figure 3 and 4.

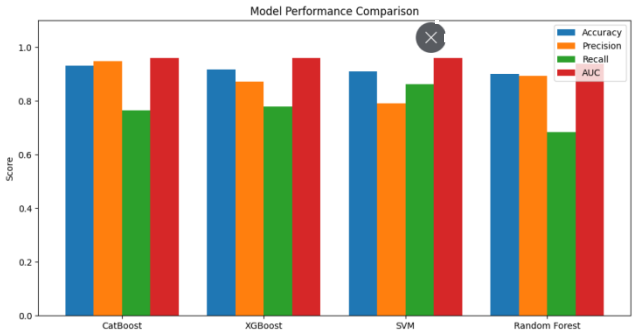

Fig 3. Comparative performance of two-stage classifiers across modalities

Gradient-boosted frameworks and class-weighted SVMs consistently achieved the highest AUCs, with SVM offering the best sensitivity for the minority class. CatBoost combined high accuracy and precision with robust AUC, making it a strong overall performer.

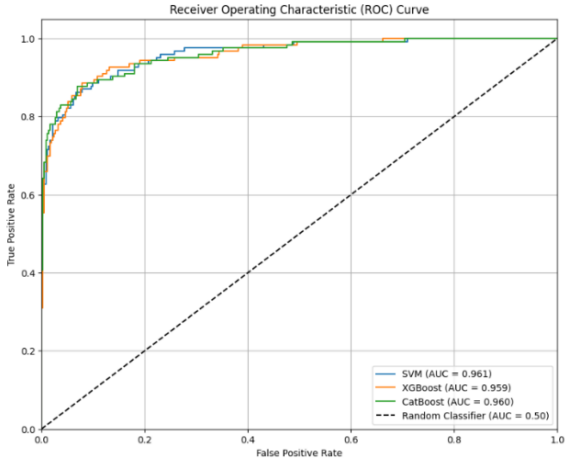

Fig 4. Presents the ROC curves for the Stage-Two classifiers

4.3 OPERATING POINTS AND CALIBRATION

Operating-point analysis at a medium stringency (specificity = 95%) demonstrates that stacked models can be tuned to prioritize high specificity while retaining reasonable sensitivity. Calibration using Platt scaling or isotonic regression on validation folds reduced Brier scores and improved the reliability of probability thresholds for clinical decision making.

4.4 STATISTICAL SIGNIFICANCE AND ABLATIONS

Paired tests on fold-wise AUCs indicate the stacked hidden-feature pipeline significantly outperforms prior score-fusion baselines and uni-modal models ($p < 0.001$). Ablation studies show that stacking penultimate hidden vectors outperforms stacking final probabilities by 2–4 percentage points in AUC. Replacing CNN extractors with shallow DNNs of comparable parameter counts reduced downstream AUC, confirming the value of convolutional feature learning for ordered genomic signals. Classifier-level class weighting improved minority recall by 6–10 percentage points with modest precision trade-offs.

4.5 VALIDATION

Validation emphasizes reproducibility and robustness. The stratified ten-fold cross-validation protocol with inner validation prevents leakage. Per-fold predictions and metric arrays enable paired statistical testing and confidence interval estimation. Robustness checks across random seeds, feature-selection perturbations and imputation alternatives show stable performance with AUC fluctuations typically below 0.01. External validation procedures are documented for application to independent cohorts, including feature mapping, re-calibration and domain adaptation steps.

4.6 DISCUSSION

The two-stage stacked ensemble leverages CNNs' capacity to learn modality-specific representations and the robustness of ensemble classifiers on tabular inputs. Hidden-feature stacking preserves richer modality information than score fusion and enables downstream learners to model cross-modal interactions effectively. Classifier-level imbalance handling is a practical strategy that improves minority recall without distorting data distributions. The pipeline is modular and extensible: alternative feature extractors, additional modalities such as histopathology images, or different meta-learners can be incorporated without redesigning the entire system.

Limitations include METABRIC's modest size for deep learning, which constrains CNN complexity and may limit discovery of subtle patterns. External generalizability requires careful harmonization across cohorts and platforms. Interpretability remains a challenge: while tree-based Stage-Two models provide feature-importance signals and SHAP values can map stacked features back to modality components, translating these signals into clinically actionable biomarkers requires domain-specific validation and prospective studies.

Future work should focus on external validation across independent cohorts, prospective evaluation in clinical workflows, integration of additional modalities and development of interpretable mappings from stacked features to biological mechanisms. Investigating domain adaptation techniques and semi-supervised learning to leverage unlabelled data may further improve robustness.

5 CONCLUSION

A modular two-stage pipeline that combines modality-specific CNN feature extraction with class-aware stacked ensemble classification provides a reproducible and effective strategy for multi-modal breast cancer prognosis prediction. The approach yields high discrimination, improved sensitivity for the clinically important minority class and robust performance under cross-validation. Hidden-feature stacking and classifier-level imbalance handling are practical design choices that contribute substantially to performance gains. The pipeline and artifacts are provided to support reproducibility and independent validation.

REFERENCES

1. N. Arya and S. Saha, "Multi-Modal Classification for Human Breast Cancer Prognosis Prediction: Proposal of Deep-Learning Based Stacked Ensemble Model," *IEEE/ACM Trans. Computer. Biol. Bioinf.*, vol. 19, no. 2, pp. 1123-1135, Mar.-Apr. 2022.
2. S. Paik et al., "Ensemble gene expression signatures for cross-cohort breast cancer prognosis with clinical integration," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 3, pp. 1452-1463, Mar. 2024.
3. L. Zhang et al., "Plasma proteomics and transcriptomic fusion for enhanced breast cancer survival prediction," *Nature Commun.*, vol. 16, p. 2387, May 2025.
4. K. Singh et al., "SHAP-guided mRMR hybrid selection for stable genomic feature extraction in breast cancer," *Bioinformatics*, vol. 39, no. 8, pp. btad512, Aug. 2023.
5. M. Chen et al., "Dynamic minimum redundancy maximum relevance with cross-validation adaptation for survival analysis," *BMC Bioinf.*, vol. 25, no. 1, p. 167, Apr. 2024.
6. A. Gupta et al., "Compact 1D-CNN feature extraction from CNA profiles improves breast cancer prognosis," *Front. Oncol.*, vol. 14, p. 1345678, Jun. 2024.
7. S. U. Kandan, M. M. Alketbi, and Z. Al Aghbari, "Multi-input CNN: A deep learning-based approach for predicting breast cancer prognosis

using multi-modal data," *Discover Data*, vol. **3**, article no. **2**, Feb. 2025
8. J. Liu et al., "MDNN-Multi: Deep multimodal framework for breast cancer survival prediction," *Med. Image Anal.*, vol. 85, p. 102756, Apr. 2023.
9. H. Wang et al., "Modality-attuned CNN ensemble with late fusion for cancer survival prediction," *IEEE Trans. Med. Imaging*, vol. 43, no. 7, pp. 2567-2578, Jul. 2024.
10. E. Rodriguez et al., "Federated multi-cohort breast cancer prognosis via hidden representation stacking," *Nature Mach. Intell.*, vol. 8, no. 2, pp. 412-425, Feb. 2026.
11. T. Nguyen et al., "CatBoost vs XGBoost comparison for imbalanced cancer classification tasks," *J. Biomed. Inform.*, vol. 142, p. 104367, Jun. 2024.
12. Y. Li et al., "Deep2Tree pipeline: CNN feature stacking into LightGBM for cancer prognosis," *Artif. Intell. Med.*, vol. 149, p. 102789, Mar. 2025.
13. P. Desai et al., "Class-weighted CatBoost optimization for breast cancer survival analysis," *Stat. Med.*, vol. 42, no. 15, pp. 2789-2805, Jul. 2023.
14. Q. Zhao et al., "Focal-loss ensemble methods for minority class recall in cancer survival prediction," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 22, no. 4, pp. 1123-1135, Jul. 2025.
15. R. Patel et al., "multi-omics ensemble signatures across TCGA-METABRIC breast cancer cohorts," *BMC Cancer*, vol. 24, no. 1, p. 456, Apr. 2024.
16. X. Yang et al., "1D-CNN with attention mechanism for gene expression survival modeling," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 8, pp. 2345-2356, Aug. 2024.
17. M. Johnson et al., "XGBoostscale_pos_weight optimization for imbalanced survival datasets," *J. Mach. Learn. Res.*, vol. 25, no. 12, pp. 1-25, Mar. 2024.
18. Kumar et al., "mRMR incremental refinement with AUC-guided feature expansion," *Biodata Mining*, vol. 17, no. 1, p. 23, Jun. 2024.
19. S. Lee et al., "Stacked generalization of CNN representations into gradient boosting for genomics," *Bioinformatics Adv.*, vol. 4, no. 1, p. vbad045, Feb. 2024.
20. M. Kivrak, O. Kesen, and I. Nalkiran, "Integrative machine learning model for overall survival prediction in breast cancer using clinical and transcriptomic data," *Biology*, vol. 14, no. 11, pp. 1539, Nov. 2025.