GMR
The Original

BAYESIAN-OPTIMISED TABULAR ATTENTION LEARNING FOR FOLLOW-UP-INFORMED HEART FAILURE MORTALITY PREDICTION

Sonal Ayyappan[1] , Naresh Sharma[2] , M Mohammed Mustafa[3] , M Pyingkodi[4] , Suresh H[5] , Kavitha V K[6] , Gourav Kalra[7] , Reji R[8]

[1]Associate Professor , SCMS School of Engineering and Technology, Kerala, India
[2]Assistant Professor, Faculty of Engineering and Technology, SRM Institute of Science and Technology, DELHI-NCR Campus,
Delhi-Meerut Road, Modinagar, Ghaziabad, Uttar Pradesh, India.
[3]Associate Professor, Sri Eshwar College of Engineering, Tamil Nadu, India
[4]Associate Professor , Kongu Engineering College, Perundurai , Tamil Nadu, India
[5]Assistant Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India
[6]Associate Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India
[7]Associate Professor, School of Engineering and Technology, CGC University, Punjab, India
[8]Professor, Thangal Kunju Musaliar Institute of Technology, Kerala, India
Corresponding author
E-mail: divyasal009@cearanmula.ac.in (DS) sonal@scmsgroup.org[1] , nrssharma@gmail.com[2] , mohammedmustafa.m@sece.ac.in[3], pyingkodikongu@gmail.com[4] , sureshh@tkmit.ac.in[5] , kavithaniji123@gmail.com[6] , gkalra89@gmail.com[7] , rejir@tkmit.ac.in[8]

Abstract
Heart failure mortality prediction remains challenging because routinely available clinical variables capture heterogeneous and interaction-dependent risk patterns. This study proposes a Bayesian-optimised deep neural framework for follow-up-informed mortality prediction using structured heart failure records. Five compact neural architectures were evaluated, including multilayer, residual, wide-and-deep, autoencoder-based, and lightweight tabular attention models. Hyperparameters were optimised using Optuna with a Tree-structured Parzen Estimator strategy, followed by repeated cross-validation, pooled out-of-fold evaluation, hold-out validation, calibration assessment, and diagnostic visualisation. The UCI Heart Failure Clinical Records dataset was used, with 299 patient records and death_event as the target variable. The proposed TabAttentionLite model achieved the strongest overall performance, with repeated cross-validation ROC-AUC of 0.8930 and pooled out-of-fold ROC-AUC of 0.8898. Hold-out validation produced ROC-AUC of 0.8570, indicating useful discrimination on unseen samples from the same dataset. These findings suggest that lightweight attention-based feature interaction modelling can improve follow-up-informed mortality-risk discrimination compared with the evaluated deep neural alternatives. Because follow-up time was included as a predictor, the model should be interpreted as a follow-up-informed prediction framework rather than a baseline-only clinical decision model.

Keywords Heart failure; mortality prediction; tabular attention; deep learning; Bayesian optimisation.

1. INTRODUCTION

Heart failure is a major cardiovascular syndrome with substantial global and United Kingdom public health relevance. Globally, heart failure affects more than 64 million people and is associated with high mortality, recurrent hospitalisation, impaired functional capacity, and considerable healthcare expenditure [1]. In the UK, more than one million people are currently living with heart failure, making it a major contributor to cardiovascular morbidity and long-term NHS care demand [2]. The wider cardiovascular burden is also projected to increase internationally; global cardiovascular prevalence is expected to rise substantially between 2025 and 2050, with cardiovascular deaths projected to reach 35.6 million by 2050 [3]. These statistics indicate the continuing need for more effective data-driven strategies for risk assessment, monitoring, and prognosis modelling in heart failure.

The clinical burden of heart failure is intensified by ageing populations, multimorbidity, delayed diagnosis, and repeated hospital admissions. In the UK, cardiovascular services continue to face pressure from growing demand for diagnostic testing, treatment, and long-term management [4]. Heart failure prognosis is particularly difficult because patients differ widely in age, renal function, cardiac function, serum biomarkers, comorbidity profile, treatment history, and follow-up duration. As a result, two patients with similar baseline presentations may have markedly different mortality risk trajectories. This heterogeneity makes heart failure an important target for machine-learning models that can combine multiple routinely collected clinical variables into patient-specific risk estimates.

Existing machine-learning studies in heart failure frequently formulate prognosis as binary mortality prediction [5]. In this setting, death-event status is predicted without explicitly considering follow-up-informed risk structure or the difference between baseline prediction and longitudinal outcome observation. Although binary classification is simple to implement and allows direct reporting of accuracy, F1-score, and ROC-AUC, it can oversimplify the clinical interpretation of heart failure prognosis. This is particularly important when follow-up time is included as a predictor, because the task becomes follow-up-informed mortality prediction rather than baseline-only risk prediction. Clear separation between these two formulations is necessary for scientifically valid model interpretation.

Structured clinical datasets such as the UCI Heart Failure Clinical Records dataset provide a compact benchmark for methodological evaluation. Although small, this dataset remains widely used for heart failure machine-learning studies because it provides a reproducible public benchmark. However, the limited sample size means that performance estimates based on a single train–test split can be unstable; repeated cross-validation, out-of-fold evaluation, confidence intervals, calibration assessment, and hold-out validation are therefore necessary to support stronger methodological claims.

Deep learning for tabular clinical data remains challenging because structured datasets are often small, heterogeneous, and sensitive to preprocessing and hyperparameter selection [6] [7]. Unlike image or signal tasks, tabular clinical prediction may not always benefit from very deep architectures. Compact neural networks, residual multilayer perceptrons, autoencoder-based classifiers, wide-and-deep networks, and lightweight tabular attention models may be more appropriate for small structured datasets. These architectures can model nonlinear feature interactions while limiting excessive model complexity. Nevertheless, their performance depends strongly on hyperparameters such as hidden dimension, depth, dropout, learning rate, weight decay, latent dimension, and training schedule.

Bayesian hyperparameter optimisation provides a systematic strategy for improving neural model selection in small clinical datasets. Optuna with a Tree-structured Parzen Estimator can efficiently search complex hyperparameter spaces without relying on manual trial-and-error tuning. For clinical machine-learning studies, optimisation should be combined with robust validation rather than reported only as a best single validation score. A stronger experimental design should include repeated cross-validation, pooled out-of-fold prediction, hold-out testing, calibration analysis, confusion matrix assessment, precision–recall analysis, and confidence intervals.

This study develops a Bayesian-optimised deep neural comparison framework for follow-up-informed heart failure mortality prediction using structured clinical records. Five neural architectures were evaluated: DeepMLP, ResidualMLP, WideDeepMLP, autoencoder-based classification, and TabAttentionLite. Each model was optimised using Optuna/TPE and validated using repeated 5-fold cross-validation with five repeats, pooled out-of-fold evaluation, and hold-out testing. The best-performing model, TabAttentionLite, achieved a repeated cross-validated ROC-AUC of 0.8930 with a 95% confidence interval of 0.8770–0.9095, pooled out-of-fold ROC-AUC of 0.8898, and hold-out ROC-AUC of 0.8570. The main contributions of this study are: (i) a deep-only comparative framework for heart failure mortality prediction; (ii) Bayesian optimisation of five neural architectures for structured clinical data; (iii) repeated cross-validation with confidence-interval reporting; (iv) pooled out-of-fold and hold-out validation; and (v) generation of paper-ready validation artefacts, including dataset sample visualisation, class-distribution analysis, correlation mapping, ROC curves, precision–recall curves, calibration plots, confusion matrices, and validation sample predictions.

The remainder of this paper is structured as follows. Section 2 reviews related work on heart failure mortality prediction, deep learning for structured clinical data, tabular attention models, Bayesian hyperparameter optimisation, and validation practices in clinical machine learning. Section 3 describes the materials and methods, including the dataset, feature formulation, preprocessing pipeline, neural architectures, optimisation strategy, and experimental protocol. Section 4 presents the results, covering model comparison, repeated cross-validation, pooled out-of-fold evaluation, hold-out validation, calibration assessment, and diagnostic visualisations. Section 5 discusses the implications of follow-up-informed prediction, the performance of tabular attention learning, small-sample modelling limitations, and the distinction between augmented and baseline-only prediction settings. Section 6 concludes the paper and outlines future work involving external validation, larger multi-centre cohorts, and prospective clinical evaluation.

2. LITERATURE REVIEW

Heart failure mortality prediction has been widely investigated using statistical learning, conventional machine-learning models, and more recently deep learning methods applied to structured clinical records. Existing studies commonly report useful discrimination for mortality or readmission prediction; however, the evidence remains inconsistent because of small datasets, heterogeneous feature sets, imbalanced outcomes, limited external validation, and frequent reliance on single train–test splits. In addition, many studies frame heart failure prognosis as binary mortality classification without clearly distinguishing baseline-only prediction from follow-up-informed prediction, particularly when follow-up time is included as an input variable. Deep learning approaches for tabular heart failure data also remain less explored than traditional models, and reported results often lack systematic

hyperparameter optimisation, pooled out-of-fold validation, calibration analysis, confidence intervals, and diagnostic visualisation. These limitations indicate the need for a rigorously validated deep learning framework that combines neural architecture comparison, Bayesian optimisation, repeated cross-validation, hold-out assessment, and transparent reporting of follow-up-informed mortality prediction.

Heart failure is clinically heterogeneous, and this heterogeneity has direct implications for prediction modelling. Patients may present with reduced ejection fraction, mildly reduced ejection fraction, preserved ejection fraction, acute or decompensated heart failure, or advanced disease. These clinical forms differ in underlying mechanisms, comorbidity patterns, treatment response, and prognosis. For example, mortality risk may be influenced by impaired systolic function, renal dysfunction, electrolyte imbalance, age-related frailty, diabetes, hypertension, anaemia, and smoking status. Therefore, mortality prediction from structured records requires models capable of capturing nonlinear and interaction-dependent relationships among routinely collected clinical variables.

Conventional machine-learning models remain dominant in heart failure prediction studies [8] [9] [10]. Logistic regression, support vector machines, random forests, gradient boosting, and XGBoost have been widely used because they perform well on tabular clinical datasets and are relatively straightforward to train. These models often achieve competitive discrimination, particularly when the number of features is limited and the dataset size is small. However, this dominance also reveals a methodological gap: fewer studies have systematically evaluated whether compact deep neural architectures can provide comparable or improved performance for structured heart failure prediction. Moreover, many existing studies prioritise accuracy or ROC-AUC alone, with limited attention to calibration, class imbalance, confidence intervals, or pooled out-of-fold validation.

Deep learning offers a flexible approach for modelling nonlinear feature relationships in clinical data, but its use in compact tabular datasets requires careful design [11] [12] [13]. Unlike imaging or signal-based tasks, structured heart failure datasets usually contain a small number of continuous and binary variables. Very deep architectures may therefore overfit, while compact neural models may be more suitable. Multilayer perceptrons can model nonlinear relationships, residual neural networks can support deeper feature transformation, wide-and-deep networks can combine direct and learned representations, and autoencoder-based classifiers can extract lower-dimensional latent features. However, the comparative behaviour of these architectures in heart failure mortality prediction remains insufficiently examined under consistent optimisation and validation conditions.

Tabular attention learning has emerged as a promising direction for structured clinical prediction [14] [15]. Attention-based models can learn feature-level interactions and may be useful where mortality risk depends on combined patterns across demographic, laboratory, cardiac-function, and follow-up variables. For heart failure data, interactions between age, ejection fraction, serum creatinine, serum sodium, comorbidity indicators, and follow-up time may contain clinically relevant risk information. A lightweight tabular attention model is particularly suitable for small structured datasets because it avoids the complexity of large transformer architectures while retaining the ability to model feature interactions. Nevertheless, tabular attention remains less frequently evaluated than traditional machine-learning models in compact heart failure benchmarks.

Bayesian hyperparameter optimisation is important for neural clinical prediction because deep model performance depends strongly on architecture and training choices [16] [17]. Hidden dimension, network depth, dropout rate, learning rate, weight decay, latent dimension, reconstruction loss weight, and training duration can all substantially affect model behaviour. Manual tuning may produce unstable or selectively favourable results. Optuna with Tree-structured Parzen Estimator optimisation provides a systematic strategy for searching neural hyperparameter spaces and selecting model configurations. However, optimisation results should not be reported in isolation. A high validation score from a single optimisation split may not generalise, particularly in small datasets; therefore, repeated cross-validation and hold-out testing are required to assess robustness.

Validation practice remains a major limitation in heart failure machine-learning studies [18] [19]. Single train–test splits can produce unstable estimates, especially in datasets with fewer than several hundred patients. Repeated cross-validation provides a more reliable estimate of performance variation, while pooled out-of-fold evaluation allows patient-level prediction assessment across all validation folds. Hold-out testing provides an additional independent demonstration, although it remains limited when drawn from the same public dataset. Calibration curves and Brier scores are also important because a model with strong discrimination may still produce poorly calibrated risk probabilities. Precision–recall analysis is relevant because death events are often the minority class, making ROC-AUC alone insufficient for evaluating clinical usefulness.

Another important issue is the distinction between baseline-only prediction and follow-up-informed prediction [20]. If follow-up time is excluded from the predictors, the model estimates mortality risk using only baseline clinical variables. If follow-up time is included, the model becomes follow-up-informed because it uses information related to the observation period. This distinction is essential for scientific reporting. Follow-up-informed prediction can be methodologically useful and may improve discrimination, but it should not be described as admission-time or baseline-only clinical prediction. Clear task definition improves interpretability and reduces the risk of overstated clinical claims.

To address these limitations, the present study proposes a Bayesian-optimised deep neural comparison framework for follow-up-informed heart failure mortality prediction. Five neural architectures are evaluated: DeepMLP,

ResidualMLP, WideDeepMLP, autoencoder-based classification, and TabAttentionLite. Each model is optimised using Optuna/TPE and assessed using repeated five-fold cross-validation with five repeats, pooled out-of-fold evaluation, independent hold-out validation, calibration analysis, and diagnostic visualisation. This design addresses key gaps in the literature by providing a deep-only comparison of tabular neural models, systematic hyperparameter optimisation, confidence-supported validation, and explicit framing of the augmented feature setting as follow-up-informed mortality prediction rather than baseline-only clinical risk estimation.

3. MATERIALS AND METHODS

This section describes the methodological framework used to develop and evaluate a Bayesian-optimised deep learning approach for follow-up-informed heart failure mortality prediction. The study was designed as a structured clinical prediction experiment using a publicly available benchmark dataset. The methodology consisted of four main stages: dataset formulation and preprocessing, development of five compact neural architectures for tabular clinical data, Bayesian hyperparameter optimisation with repeated validation, and performance assessment using discrimination, calibration, confidence-interval estimation, and diagnostic visualisation. The proposed TabAttentionLite architecture is shown in Fig. 1.

3.1 Dataset and Study Design

The UCI Heart Failure Clinical Records dataset was used as the benchmark cohort. The dataset contains 299 patient records with demographic, clinical, laboratory, follow-up, and outcome variables. The binary target variable was death_event, where a value of 1 indicates occurrence of death during the recorded follow-up period and 0 indicates survival during the observation period. The study used the augmented feature setting, in which follow-up time was included as an input variable together with baseline clinical variables. Therefore, the task was defined as follow-up-informed mortality prediction rather than baseline-only clinical prediction.

Twelve predictors were used: age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelet count, serum creatinine, serum sodium, sex, smoking status, and follow-up time. The inclusion of follow-up time was explicitly treated as an augmented modelling condition because this variable contains information related to the observation period. This distinction was maintained throughout the study to avoid presenting the model as an admission-time or baseline-only mortality prediction system.

The dataset was represented as:

$D = \{(x_i, y_i)\}_{i=1}^{N}$ where $N = 299$, $x_i \in \mathbb{R}^d$ denotes the feature vector of patient i, and $y_i \in \{0,1\}$ denotes the corresponding binary mortality label. The model estimated the probability of death event as: $\hat{y}_i = f\theta(x_i)$ where $f\theta$ denotes a neural model with trainable parameters θ, and $\hat{y}_i \in [0,1]$ denotes the predicted mortality probability for patient i. Missing values were handled using median imputation, and numerical variables were standardised using z-score normalisation:

$$z_{ij} = (x_{ij} - \mu_j) / \sigma_j \quad (1)$$

where x_{ij} is the original value of feature j for patient i, μ_j is the mean of feature j estimated from the training subset, and σ_j is the corresponding standard deviation. To prevent information leakage, imputation and standardisation parameters were fitted only on the training subset within each validation fold and were then applied to the corresponding validation or hold-out subset.

3.2 Proposed Deep Learning Framework

The proposed framework compared five compact neural architectures under a unified optimisation and validation protocol: DeepMLP, ResidualMLP, WideDeepMLP, AEClassifier, and TabAttentionLite. These architectures were selected to represent complementary deep learning strategies for structured clinical data, including direct nonlinear modelling, residual representation learning, combined shallow–deep representation, latent representation learning, and attention-based feature interaction modelling.

DeepMLP consisted of stacked fully connected layers with batch normalisation, rectified linear unit activation, and dropout. ResidualMLP extended this structure using residual blocks to support deeper nonlinear transformation while reducing degradation during training. WideDeepMLP combined the original standardised input vector with a learned deep representation before final classification, allowing both direct and transformed feature information to contribute to the prediction. AEClassifier used an autoencoder module to learn a compact latent representation before supervised mortality classification. TabAttentionLite projected tabular variables into feature-level embeddings and applied a lightweight attention mechanism to capture interactions among clinical predictors.

For the multilayer neural models, the hidden representation at layer l was defined as:

$$h^l = \phi(W^l h^{l-1} + b^l) \quad (2)$$

where h^l is the hidden representation at layer l, W^l and b^l are trainable weights and bias terms, and $\phi(.)$ denotes the nonlinear activation function. The final mortality probability was obtained using the sigmoid function:

$$\hat{y}_i = \sigma(o_i) = 1 / (1 + e^{-oi}) \quad (3)$$

where o_i is the output logit for patient i. For AEClassifier, the input vector was encoded into a latent representation: $z_i = g\theta(x_i)$ and reconstructed as: $\tilde{x}_i = r\psi(z_i)$ where $g\theta(.)$ denotes the encoder and $r\psi(.)$ denotes the decoder. The autoencoder classifier was trained using a combined objective:

$L\_AE = L\_BCE(y_i, \hat{y}_i) + \lambda \|x_i - \tilde{x}_i\|^2$ (4)
where L_BCE denotes binary classification loss, λ controls the contribution of the reconstruction penalty, and $\|x_i - \tilde{x}_i\|^2$ measures reconstruction error.
For TabAttentionLite, feature-level embeddings were processed using a lightweight attention operation:
$\text{Attention}(Q, K, V) = \text{softmax}(QK^T / \sqrt{d_k})V$ (5)
where Q, K, and V denote query, key, and value matrices derived from embedded clinical variables, and d_k denotes the key dimension. The attention-informed representation was passed to a classifier head to estimate mortality probability.

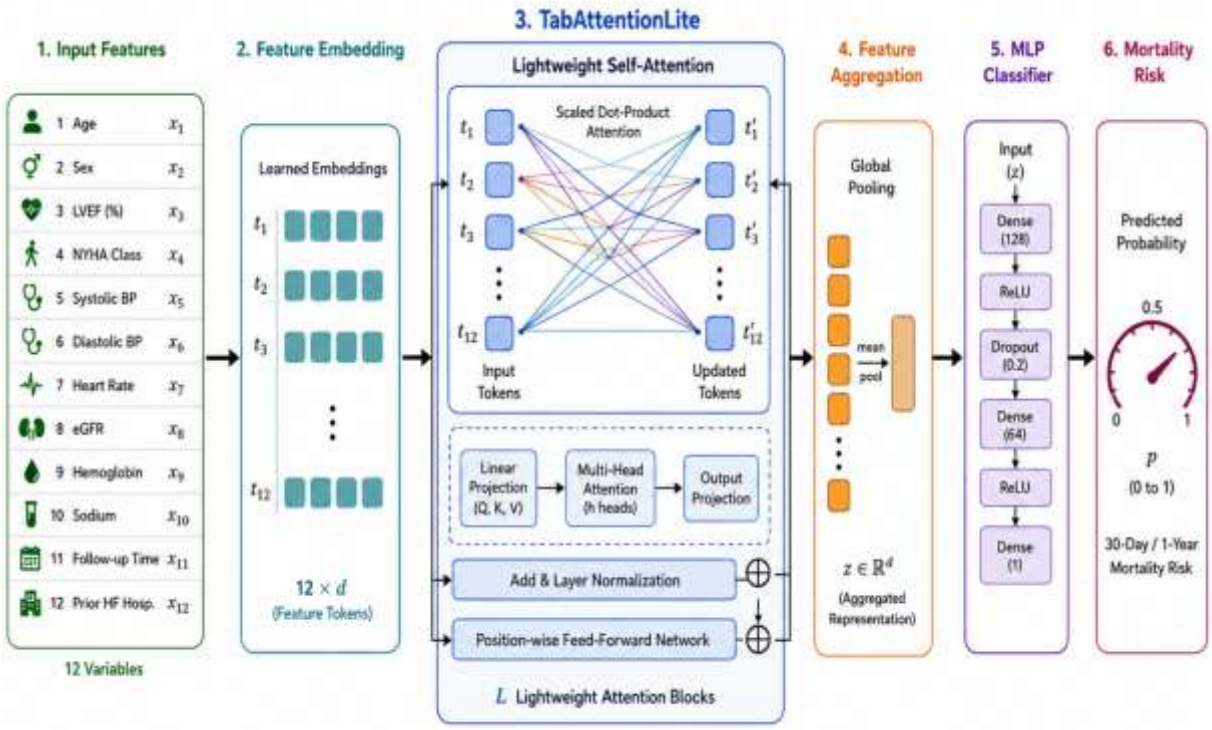

Fig. 1. Proposed TabAttentionLite architecture for follow-up-informed heart failure mortality prediction.

The model converts structured clinical inputs into feature embeddings, applies lightweight attention to learn feature-level interactions, aggregates the learned representation, and estimates mortality risk using an MLP classifier.

3.3 Optimisation, Training and Validation

Hyperparameter optimisation was performed separately for each neural architecture using Optuna with the Tree-structured Parzen Estimator strategy. Architecture-specific optimisation was used to avoid favouring any model through manually selected default settings. The search space included hidden dimension, network depth, number of residual blocks, dropout rate, learning rate, weight decay, latent dimension, reconstruction loss weight, embedding dimension, and number of attention heads where applicable.

The optimisation objective combined discrimination and classification quality. ROC-AUC was prioritised as the primary ranking metric, while F1-score was included to reflect event-class detection under class imbalance. The optimisation score was defined as:

$S = 0.75 \times \text{ROC-AUC} + 0.25 \times \text{F1}$ (6)

Models were trained using weighted binary cross-entropy with logits. The weighted binary cross-entropy loss was defined as:

$L\_BCE = -(1/N) \Sigma_{i=1}^{N} [w\, y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$ (7)

where w denotes the positive-class weight estimated from the training subset. AdamW optimisation was used for parameter updates, and dropout was applied for regularisation. Early stopping was based on validation ROC-AUC, and the model state with the highest validation performance was retained.

Model performance was evaluated using repeated stratified five-fold cross-validation with five repeats. Stratification preserved the proportion of death-event and non-death-event cases across folds. Within each fold, preprocessing was fitted only on the training data, the neural model was trained using an internal validation split, and predictions were generated for the held-out fold.

To reduce stochastic variation caused by neural weight initialisation and small sample size, each architecture was evaluated using multiple independently trained neural instances. The final patient-level probability was obtained by averaging ensemble predictions:

$\hat{p}_i = (1/M) \Sigma_{m=1}^{M} \hat{p}_{im}$ (8)

where M denotes the number of ensemble members and $\hat{p}_{im}$ denotes the predicted probability from the m-th neural instance.

Pooled out-of-fold evaluation was then performed by combining predictions from all validation folds. This provided patient-level prediction assessment across the repeated validation process. A classification threshold was selected to optimise F1-score using pooled out-of-fold predictions. A separate stratified hold-out evaluation was also conducted for the best-performing model to provide an additional assessment of generalisation on unseen samples drawn from the same dataset.

3.4 Evaluation Metrics and Statistical Reporting

Model performance was evaluated using discrimination, classification, and calibration-related metrics. ROC-AUC was selected as the primary discrimination metric because it measures the model's ability to rank death-event and non-death-event cases across thresholds. PR-AUC was included because the death-event class was smaller than the non-event class. Accuracy, balanced accuracy, precision, recall, and F1-score were reported to characterise threshold-dependent prediction performance.

The Brier score was used to quantify probabilistic prediction error:

Brier Score = $(1/N) \Sigma_{i=1}^{N} (\hat{y}_i - y_i)^2$ (9)

where $\hat{y}_i$ denotes the predicted mortality probability, y_i denotes the observed binary outcome, and N denotes the total number of evaluated samples. Lower Brier score values indicate better probabilistic prediction performance. Bootstrap confidence intervals were calculated for repeated cross-validation metrics to quantify uncertainty around performance estimates. Calibration curves were used to compare predicted mortality probabilities with observed event frequencies. Diagnostic visualisations included dataset sample plots, class-distribution analysis, missing-value plots, correlation heatmaps, model comparison plots, pooled ROC curves, precision–recall curves, calibration plots, hold-out confusion matrices, predicted-probability distributions, and validation sample prediction tables. All fold-level results, pooled out-of-fold predictions, hold-out predictions, hyperparameter optimisation records, confidence intervals, and publication-ready figures were exported to support reproducibility and transparent reporting.

4. RESULTS

This section presents the experimental findings obtained from the Bayesian-optimised deep neural framework for follow-up-informed heart failure mortality prediction. Five deep neural architectures were evaluated under the same optimisation and validation protocol: DeepMLP, ResidualMLP, WideDeepMLP, AEClassifier, and TabAttentionLite. The evaluation was performed using the augmented feature setting, in which follow-up time was included with baseline clinical variables. Model performance was assessed using repeated five-fold cross-validation with five repeats, pooled out-of-fold evaluation, hold-out validation, calibration analysis, and diagnostic visualisation.

4.1 Dataset Characteristics

The dataset contained 299 patient records, consisting of 203 non-death-event cases and 96 death-event cases. Twelve predictors were used in the augmented setting: age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelet count, serum creatinine, serum sodium, sex, smoking status, and follow-up time. The target variable was death_event. The class distribution showed moderate imbalance, with death-event cases representing the smaller class. Therefore, PR-AUC, recall, F1-score, balanced accuracy, and Brier score were reported in addition to ROC-AUC.

Table 1. Dataset characteristics and feature description.

Variable	Description	Data type	Role in this study
age	Age of the patient in years	Continuous	Predictor
anaemia	Presence of anaemia, coded as 0 = no and 1 = yes	Binary	Predictor
creatinine_phosphokinase	Creatinine phosphokinase level in blood	Continuous	Predictor
diabetes	Diabetes status, coded as 0 = no and 1 = yes	Binary	Predictor
ejection_fraction	Percentage of blood leaving the heart at each contraction	Continuous	Predictor
high_blood_pressure	High blood pressure status, coded as 0 = no and 1 = yes	Binary	Predictor
platelets	Platelet count in blood	Continuous	Predictor
serum_creatinine	Serum creatinine level in blood	Continuous	Predictor

Variable	Description	Data type	Role in this study
serum_sodium	Serum sodium level in blood	Continuous	Predictor
sex	Patient sex, coded as 0 = female and 1 = male	Binary	Predictor
smoking	Smoking status, coded as 0 = no and 1 = yes	Binary	Predictor
time	Follow-up duration	Continuous	Predictor in augmented follow-up-informed setting
death_event	Mortality outcome during follow-up, coded as 0 = survived and 1 = death event	Binary	Target

Table 1 summarises the variables used in this study, including their clinical meaning, data type, and role in the prediction task. It defines the augmented feature formulation by identifying the 12 predictor variables and clearly distinguishing them from the target variable, death_event.

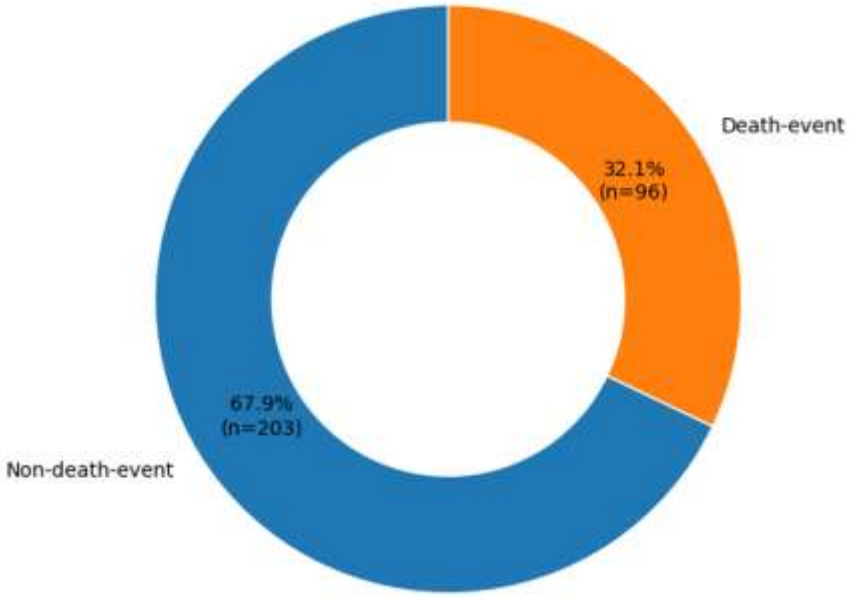

Fig. 2. Class distribution of death-event and non-death-event cases.

This figure shows the number of patients in each outcome class. It supports the use of class-sensitive evaluation metrics because the death-event class was smaller than the non-death-event class.

4.2 Comparison of Deep Neural Architectures

Five Bayesian-optimised neural architectures were compared to assess different approaches to tabular clinical prediction. DeepMLP represented direct nonlinear modelling using fully connected layers. ResidualMLP introduced residual connections to support deeper feature transformation. WideDeepMLP combined original input features with learned representations. AEClassifier used autoencoder-based latent representation learning before classification. TabAttentionLite used lightweight feature-level attention to model interactions among structured clinical predictors.

Table 2. Summary of the five evaluated deep neural architectures.

Model	Architecture type	Main modelling purpose	Key components	Output
DeepMLP	Fully connected neural network	Learns nonlinear relationships among structured clinical variables	Dense layers, batch normalisation, ReLU activation, dropout	Mortality probability
ResidualMLP	Residual multilayer perceptron	Supports deeper feature transformation while reducing training degradation	Dense input layer, residual blocks, batch normalisation, dropout	Mortality probability

Model	Architecture type	Main modelling purpose	Key components	Output
WideDeepMLP	Wide-and-deep neural network	Combines original feature information with learned nonlinear representations	Original input concatenated with deep hidden representation	Mortality probability
AEClassifier	Autoencoder-based classifier	Learns compact latent clinical representations before classification	Encoder, latent representation, decoder, classifier head	Mortality probability
TabAttentionLite	Lightweight tabular attention network	Models feature-level interactions among structured clinical predictors	Feature embeddings, attention mechanism, classifier head	Mortality probability

This table summarises the purpose and main structural component of each architecture. It provides a concise comparison of the modelling strategies used in the experiment without overloading the results section with implementation details.

Among the five models, TabAttentionLite achieved the strongest overall performance. This indicates that attention-based feature interaction modelling was more effective than direct multilayer transformation, residual transformation, wide-and-deep concatenation, or autoencoder-based representation learning for this augmented structured dataset.

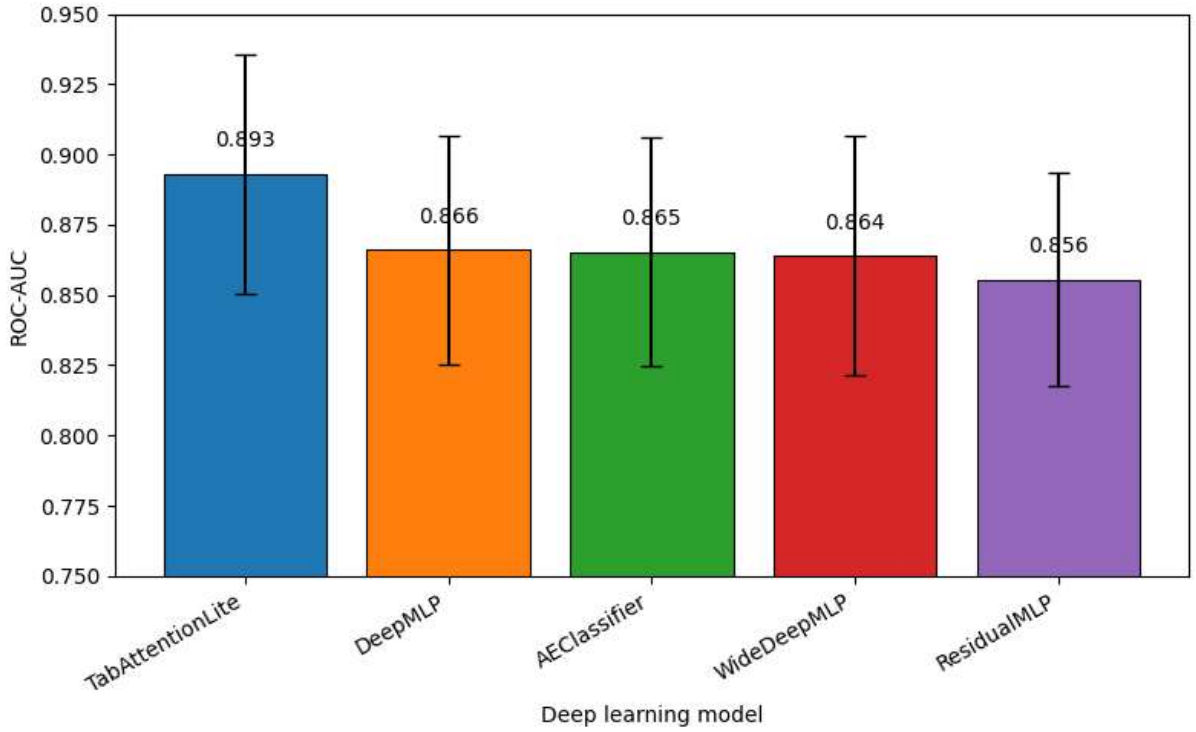

Fig. 3. Repeated cross-validation ROC-AUC comparison across the five deep models.

This figure presents the mean ROC-AUC of the five neural architectures under repeated cross-validation. It is the main model-comparison figure and visually identifies TabAttentionLite as the best-performing model.

4.3 Repeated Cross-Validation Performance

Repeated five-fold cross-validation with five repeats was used to estimate model performance and variability across validation folds. TabAttentionLite achieved the best repeated cross-validation performance, with a ROC-AUC of 0.8930 and a 95% confidence interval of 0.8770–0.9095. The model also achieved PR-AUC of 0.8103, F1-score of 0.7377, recall of 0.7622, and Brier score of 0.1261. These findings indicate that TabAttentionLite provided the strongest discrimination and probabilistic performance among the evaluated neural models.

Table 3. Repeated cross-validation performance with bootstrap confidence intervals.

Model	ROC-AUC, mean (95% CI)	PR-AUC, mean (95% CI)	F1-score, mean (95% CI)	Recall, mean (95% CI)	Brier score, mean (95% CI)
Proposed TabAttentionLite	0.8930 (0.8770–0.9095)	0.8103 (0.7793–0.8393)	0.7377 (0.7110–0.7653)	0.7622 (0.7272–0.7968)	0.1261 (0.1154–0.1369)
AEClassifier	0.8653 (0.8497–0.8812)	0.7625 (0.7306–0.7940)	0.6938 (0.6725–0.7141)	0.7601 (0.7363–0.7854)	0.1540 (0.1464–0.1616)
WideDeepMLP	0.8639 (0.8476–0.8806)	0.7638 (0.7304–0.7955)	0.6955 (0.6720–0.7201)	0.7454 (0.7168–0.7731)	0.1520 (0.1436–0.1606)
DeepMLP	0.8660 (0.8510–0.8813)	0.7712 (0.7432–0.7988)	0.7024 (0.6799–0.7242)	0.7434 (0.7085–0.7791)	0.1570 (0.1508–0.1639)
ResidualMLP	0.8556 (0.8412–0.8704)	0.7528 (0.7267–0.7780)	0.6867 (0.6659–0.7078)	0.7352 (0.7059–0.7651)	0.1587 (0.1510–0.1661)

Table 3 reports the mean repeated cross-validation performance with bootstrap 95% confidence intervals. The proposed TabAttentionLite model achieved the strongest overall repeated-CV performance, obtaining the highest ROC-AUC, PR-AUC, F1-score, and recall, together with the lowest Brier score among the five evaluated deep neural architectures. These results indicate that the proposed model provided the best balance between discrimination, event-class detection, and probabilistic prediction performance.

4.4 Pooled Out-of-Fold Evaluation

Pooled out-of-fold evaluation was performed by combining predictions generated across the repeated validation process. This analysis provided patient-level prediction assessment across all validation folds. TabAttentionLite again achieved the strongest pooled out-of-fold performance, with ROC-AUC of 0.8898, PR-AUC of 0.7886, accuracy of 0.8294, balanced accuracy of 0.8189, precision of 0.7111, recall of 0.7896, F1-score of 0.7483, and Brier score of 0.1261. The selected threshold for classification was 0.47.

Table 4. Pooled out-of-fold performance of the five deep models.

Model	**ROC-AUC**	**PR-AUC**	**Accuracy**	**Balanced accuracy**	**Precision**	**Recall**	**F1-score**	**Brier score**	**Threshold**
Proposed TabAttentionLite	**0.8898**	**0.7886**	**0.8294**	**0.8189**	**0.7111**	**0.7896**	**0.7483**	**0.1261**	0.47
DeepMLP	0.8578	0.7303	0.8094	0.7794	0.7061	0.6958	0.7009	0.1570	0.57
AEClassifier	0.8541	0.7625	0.8027	0.7890	0.6750	0.7604	0.7152	0.1540	0.48
WideDeepMLP	0.8521	0.7638	0.7993	0.7910	0.6643	**0.7813**	0.7181	0.1520	0.46
ResidualMLP	0.8432	0.7200	0.7926	0.7695	0.6701	0.7188	0.6936	0.1587	0.50

Table 4 reports pooled out-of-fold performance by aggregating validation predictions across the repeated cross-validation process. The proposed TabAttentionLite model achieved the strongest overall pooled performance, with the highest ROC-AUC, PR-AUC, accuracy, balanced accuracy, precision, recall, F1-score, and the lowest Brier score. These findings indicate that TabAttentionLite provided the best overall balance between discrimination, event-class sensitivity, classification performance, and probabilistic prediction quality among the five evaluated deep neural models.

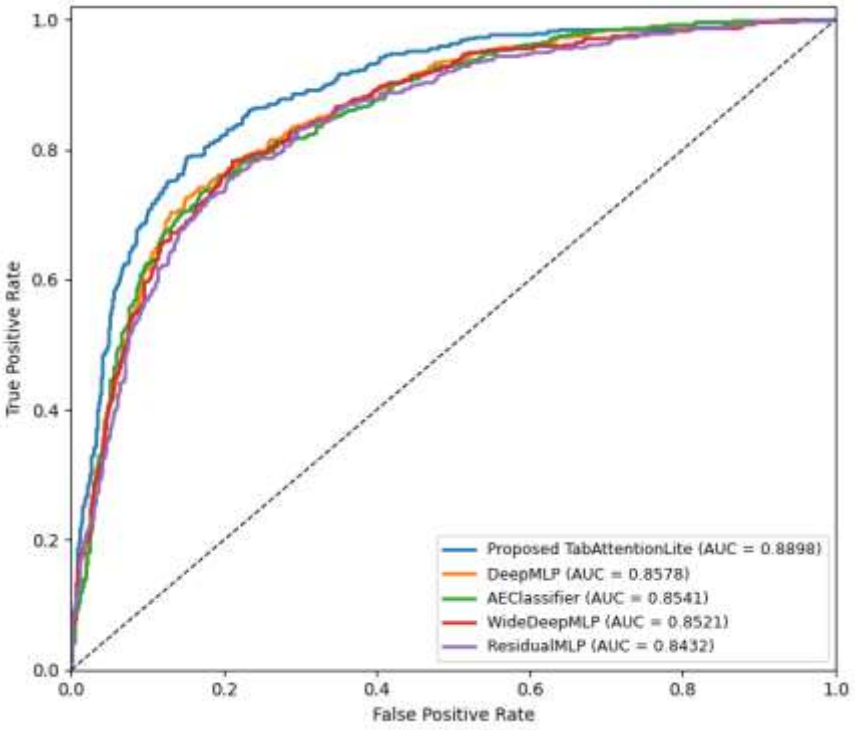

Fig. 4. Pooled out-of-fold ROC curves for the five deep models.
This figure compares the discrimination behaviour of all five models across classification thresholds. It supports the ranking observed in the repeated cross-validation results.

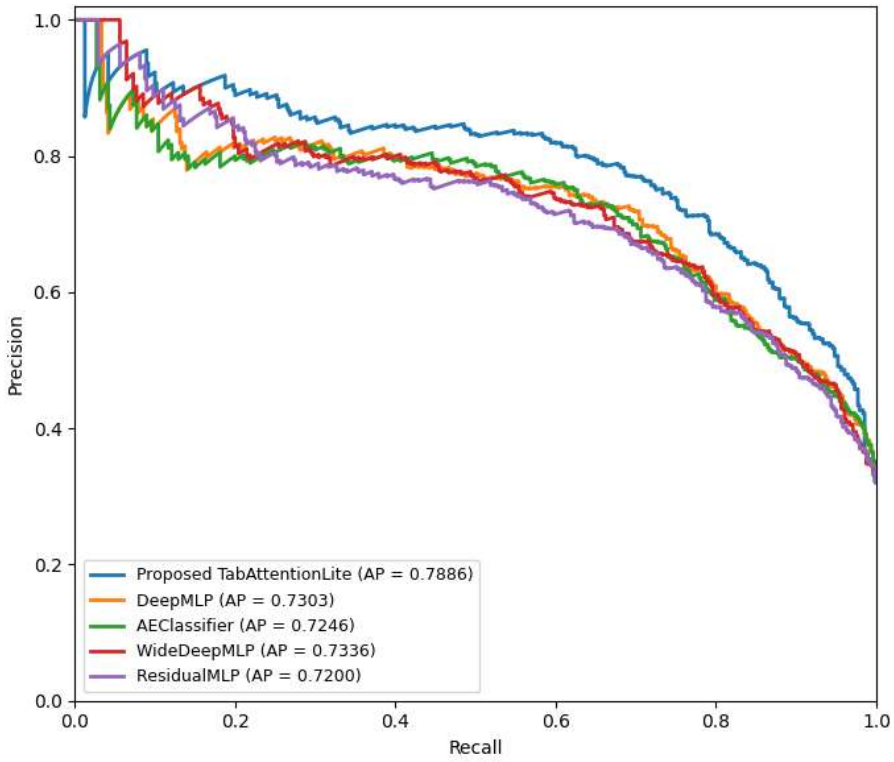

Fig. 5. Pooled out-of-fold precision–recall curves for the five deep models.

This figure evaluates event-class detection under class imbalance. It is particularly relevant because the death-event class was smaller than the non-event class.

4.5 Hold-Out Validation and Calibration

A stratified hold-out evaluation was conducted for the best-performing model, TabAttentionLite. On the hold-out subset, the model achieved ROC-AUC of 0.8570, PR-AUC of 0.7888, accuracy of 0.8267, balanced accuracy of

0.7953, precision of 0.7391, recall of 0.7083, F1-score of 0.7234, and Brier score of 0.1381. The hold-out ROC-AUC was lower than the repeated cross-validation and pooled out-of-fold values, which is expected in a small public dataset. Nevertheless, the model maintained useful discrimination on unseen samples from the same dataset.

Table 5. Hold-out validation performance of the proposed TabAttentionLite model.

Model	ROC-AUC	PR-AUC	Accuracy	Balanced accuracy	Precision	Recall	F1-score	Brier score	Threshold
Proposed TabAttentionLite	**0.8570**	**0.7888**	**0.8267**	**0.7953**	**0.7391**	**0.7083**	**0.7234**	**0.1381**	**0.47**

Table 5 reports the hold-out validation performance of the proposed TabAttentionLite model. The model achieved a hold-out ROC-AUC of 0.8570 and F1-score of 0.7234, indicating useful discrimination on unseen samples drawn from the same dataset. The hold-out performance was lower than the repeated cross-validation and pooled out-of-fold results, which is expected given the small dataset size and reinforces the need for external validation.

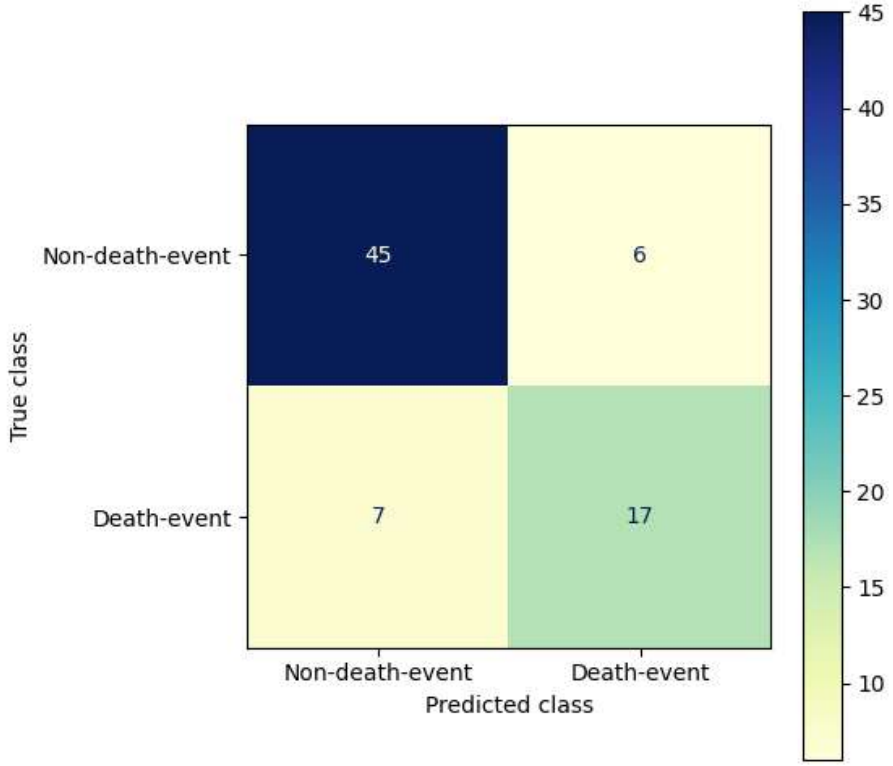

Fig. 6. Hold-out confusion matrix for TabAttentionLite.

This figure shows correct and incorrect classifications for death-event and non-death-event cases at the selected threshold. It helps interpret class-wise prediction behaviour beyond ROC-AUC.

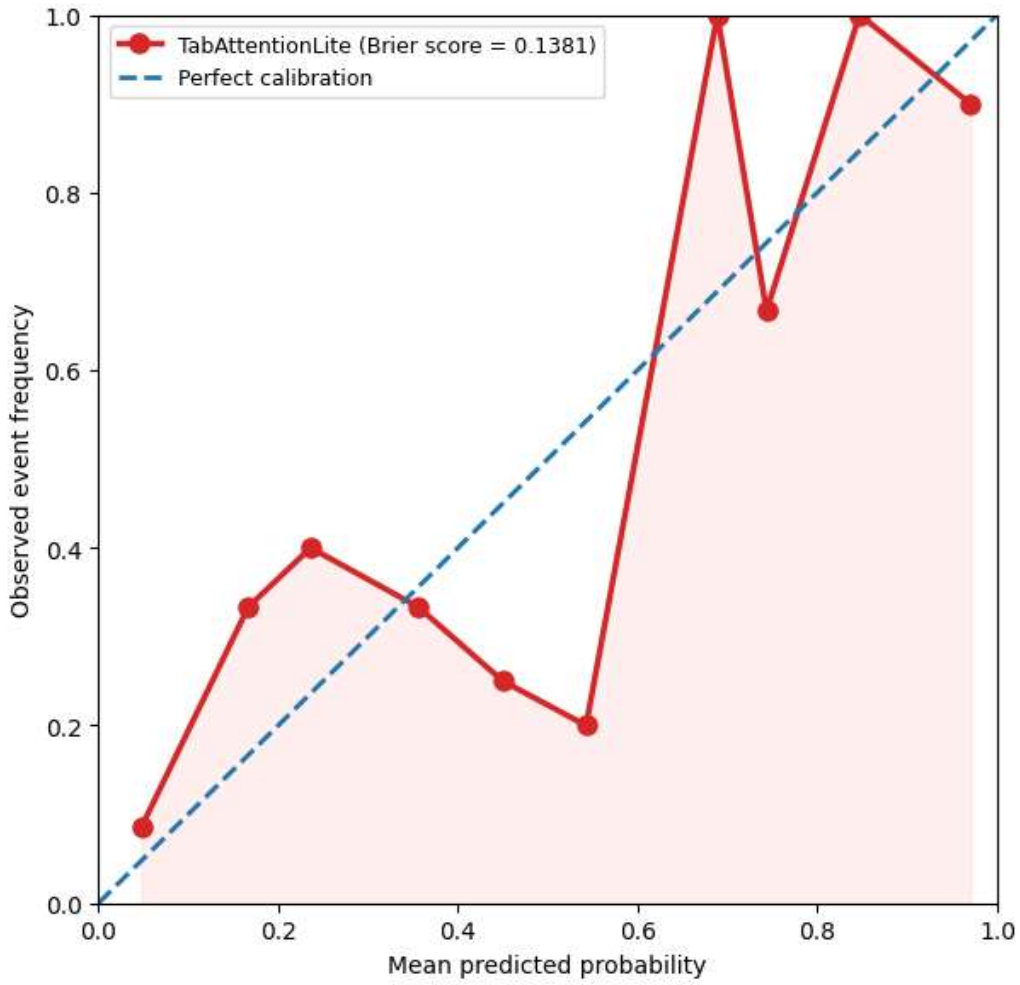

Fig. 7. Calibration curve for TabAttentionLite.

This figure compares predicted mortality probabilities with observed event frequencies. It supports the Brier score analysis and helps assess whether predicted probabilities are reasonably calibrated.

4.6 Principal Findings

The comparative evaluation identified TabAttentionLite as the strongest model among the five Bayesian-optimised deep neural architectures. In repeated five-fold cross-validation with five repeats, TabAttentionLite achieved the highest discrimination, with a ROC-AUC of 0.8930 and a 95% confidence interval of 0.8770–0.9095. The model retained comparable performance in pooled out-of-fold evaluation, achieving a ROC-AUC of 0.8898, and showed consistent but lower discrimination in hold-out testing, with a ROC-AUC of 0.8570. This performance pattern indicates that lightweight attention-based feature interaction modelling was more effective than multilayer, residual, wide-and-deep, and autoencoder-based neural alternatives for the augmented structured heart-failure dataset. The reduction between repeated cross-validation and hold-out performance also reflects the expected sensitivity of model estimates in a small public cohort. Therefore, the results support the use of Bayesian-optimised tabular attention learning for follow-up-informed mortality risk discrimination, while also showing that external validation is required before any clinical interpretation.

5. DISCUSSION

This study evaluated a Bayesian-optimised deep neural framework for follow-up-informed heart failure mortality prediction using structured clinical records. Five compact neural architectures were compared under a consistent optimisation and validation protocol: DeepMLP, ResidualMLP, WideDeepMLP, AEClassifier, and TabAttentionLite. The results showed that TabAttentionLite achieved the strongest performance across repeated cross-validation, pooled out-of-fold evaluation, and hold-out validation. This indicates that lightweight attention-based modelling of feature interactions was more effective than direct multilayer transformation, residual representation learning, wide-and-deep concatenation, or autoencoder-based latent representation for the augmented heart failure dataset.

The strongest repeated cross-validation performance was obtained by TabAttentionLite, which achieved a ROC-AUC of 0.8930 with a 95% confidence interval of 0.8770–0.9095. The model also maintained strong pooled out-of-fold performance, with ROC-AUC of 0.8898, and achieved hold-out ROC-AUC of 0.8570. The reduction from repeated cross-validation to hold-out evaluation is expected in a small public dataset and highlights the importance of reporting more than a single validation result. If only the best optimisation or cross-validation result were reported, model performance could appear more stable than it actually is. Therefore, the combined use of repeated cross-validation, pooled out-of-fold analysis, and hold-out testing provides a more balanced estimate of model behaviour.

The superior performance of TabAttentionLite suggests that feature-level interaction modelling is useful for follow-up-informed heart failure mortality prediction. Mortality risk in heart failure is unlikely to depend on one variable alone; rather, it may arise from combined patterns across age, ejection fraction, renal-function markers, serum sodium, comorbidity indicators, behavioural variables, and follow-up time. A lightweight attention mechanism can model relationships among these structured predictors without requiring a large transformer-style architecture. This may explain why TabAttentionLite outperformed the other evaluated deep models in this compact tabular dataset.

The results also show that deeper or more complex neural structures did not automatically improve performance. DeepMLP, ResidualMLP, WideDeepMLP, and AEClassifier all produced useful predictive discrimination, but they did not outperform TabAttentionLite. This is important because small structured clinical datasets are vulnerable to overfitting, and excessive architectural complexity may not translate into better generalisation. In this context, compact attention-based modelling appears to provide a stronger balance between representational capacity and regularisation.

The use of Bayesian optimisation strengthened the model-development process because neural model performance is sensitive to hyperparameter choices. Hidden dimension, network depth, dropout rate, learning rate, weight decay, latent dimension, reconstruction weight, and attention configuration can substantially influence model behaviour. By applying Optuna/TPE separately to each architecture, the study reduced the risk of favouring one model because of arbitrary parameter choices. However, optimisation performance alone was not treated as the final result. The selected configurations were further assessed through repeated validation, pooled prediction analysis, and hold-out testing, which strengthened the reliability of the comparison.

The inclusion of follow-up time requires careful interpretation. Because follow-up time was used as an input predictor, the model should be interpreted as a follow-up-informed mortality prediction framework rather than a baseline-only admission-time prediction model. This distinction is clinically and methodologically important. A baseline-only model would estimate mortality risk using information available at or near initial assessment, whereas the augmented model uses observation-period information that may improve discrimination but changes the prediction setting. Therefore, the findings should not be overstated as evidence for a purely prospective baseline clinical decision tool.

The calibration and Brier score analyses provided additional information beyond discrimination. A model may achieve strong ROC-AUC while producing poorly calibrated probabilities. In this study, the pooled out-of-fold Brier score of 0.1261 and hold-out Brier score of 0.1381 suggest that TabAttentionLite produced reasonably useful probability estimates within the constraints of the dataset. Calibration assessment is important for clinical prediction because predicted probabilities should support interpretation and risk stratification rather than merely ranking patients.

The study has some limitations. The dataset included only 299 patients, which may limit the stability and generalisability of the model. The hold-out evaluation was drawn from the same public cohort rather than an independent external dataset. In addition, follow-up time was included as a predictor, so the model should be interpreted as follow-up-informed mortality prediction rather than baseline-only clinical prediction. Although repeated cross-validation and bootstrap confidence intervals improved the reliability of evaluation, external validation using larger multi-centre cohorts remains necessary.

Despite these limitations, the study provides a focused methodological contribution. It presents a deep-only comparison of five neural architectures under a unified Bayesian optimisation and validation framework. It also reports repeated cross-validation, pooled out-of-fold evaluation, hold-out testing, calibration analysis, and diagnostic visualisation rather than relying only on a single metric or single split. This design provides a more transparent assessment of deep learning performance in a small structured heart failure dataset and demonstrates the potential of lightweight tabular attention learning for follow-up-informed mortality-risk discrimination.

6. CONCLUSION AND FUTURE WORK

This study presented a Bayesian-optimised deep neural framework for follow-up-informed heart failure mortality prediction using structured clinical records. Five compact neural architectures were evaluated under a unified validation protocol, and TabAttentionLite achieved the strongest overall performance, with repeated cross-validation ROC-AUC of 0.8930, pooled out-of-fold ROC-AUC of 0.8898, and hold-out ROC-AUC of 0.8570.

These findings suggest that lightweight attention-based modelling of feature interactions can provide effective mortality-risk discrimination in the augmented heart failure prediction setting.
The results should be interpreted as methodological evidence rather than clinical deployment readiness because the dataset contained only 299 patients and follow-up time was included as a predictor. Future work should validate the framework using larger multi-centre cohorts, independent external test sets, and baseline-only prediction settings excluding follow-up time. Further extensions should investigate survival-specific modelling, longitudinal clinical variables, multimodal data integration, calibration refinement, and explainability methods to support clinical interpretation.

DECLARATIONS

Ethics statement
This study was conducted using publicly available datasets. No research involving identifiable human participants was carried out. Accordingly, ethical approval was not required.
Consent to participate
Not applicable.
Consent to publish
Not applicable.
Funding
No funding was received for this study.
Data Availability
The dataset analysed in this study is publicly available from the UCI Machine Learning Repository under the record titled Heart Failure Clinical Records:
https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records
The dataset contains structured clinical records for 299 patients, including demographic, clinical, laboratory, follow-up, and mortality outcome variables. The data were used in accordance with the repository terms and conditions.
Author-generated code supporting the findings of this study is available in the project repository:
https://github.com/divsal009/Heart/

REFERENCES
[1] V. Valente et al., “The global epidemiology of heart failure: a comprehensive and contemporary review,” European Journal of Heart Failure, p. xuag121, Apr. 2026, doi: 10.1093/ejhf/xuag121.
[2] N. Conrad et al., “Trends in cardiovascular disease incidence among 22 million people in the UK over 20 years: population based study,” BMJ, vol. 385, p. e078523, Jun. 2024, doi: 10.1136/bmj-2023-078523.
[3] B. Chong et al., “Global burden of cardiovascular diseases: projections from 2025 to 2050,” European Journal of Preventive Cardiology, vol. 32, no. 11, pp. 1001–1015, Aug. 2025, doi: 10.1093/eurjpc/zwae281.
[4] L. Sperling et al., “WHF Roadmap for Integrated Care in People Living with – or at Risk of – Cardiovascular Disease and Multiple Long-Term Conditions,” Global Heart, vol. 21, no. 1, p. 28, Mar. 2026, doi: 10.5334/gh.1541.
[5] E. Kokori et al., “Machine learning in predicting heart failure survival: a review of current models and future prospects,” Heart Fail Rev, vol. 30, no. 2, pp. 431–442, Dec. 2024, doi: 10.1007/s10741-024-10474-y.
[6] S. Divya, L. Padma Suresh, and A. John, “Hybrid Optimization Algorithm-Based Generative Adversarial Network for Change Detection Using Pre-Operative and Post-Operative MRI,” Int. J. Patt. Recogn. Artif. Intell., vol. 36, no. 07, p. 2251007, Jun. 2022, doi: 10.1142/S0218001422510077.
[7] D. S, L. Padma Suresh, and A. John, “A Deep Transfer Learning framework for Multi Class Brain Tumor Classification using MRI,” in 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India: IEEE, Dec. 2020, pp. 283–290. doi: 10.1109/ICACCCN51052.2020.9362908.
[8] H. A. Al-Alshaikh et al., “Comprehensive evaluation and performance analysis of machine learning in heart disease prediction,” Sci Rep, vol. 14, no. 1, p. 7819, Apr. 2024, doi: 10.1038/s41598-024-58489-7.
[9] S. Shin et al., “Machine Learning vs. Conventional Statistical Models for Predicting Heart Failure Readmission and Mortality,” ESC Heart Failure, vol. 8, no. 1, pp. 106–115, Feb. 2021, doi: 10.1002/ehf2.13073.
[10] G. Bazoukis et al., “Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic review,” Heart Fail Rev, vol. 26, no. 1, pp. 23–34, Jan. 2021, doi: 10.1007/s10741-020-10007-3.
[11] C. Zhou et al., “A comprehensive review of deep learning-based models for heart disease prediction,” Artif Intell Rev, vol. 57, no. 10, p. 263, Aug. 2024, doi: 10.1007/s10462-024-10899-9.

[12] D. Li, J. Fu, J. Zhao, J. Qin, and L. Zhang, “A deep learning system for heart failure mortality prediction,” PLoS ONE, vol. 18, no. 2, p. e0276835, Feb. 2023, doi: 10.1371/journal.pone.0276835.
[13] F. Goretti, B. Oronti, M. Milli, and E. Iadanza, “Deep Learning for Predicting Congestive Heart Failure,” Electronics, vol. 11, no. 23, p. 3996, Dec. 2022, doi: 10.3390/electronics11233996.
[14] M. H. Alshayeji and S. Abed, “Heart disease prediction by tabular modeling with deep learning network and interpretability,” Mach. Learn.: Sci. Technol., vol. 6, no. 3, p. 035043, Sep. 2025, doi: 10.1088/2632-2153/adfd39.
[15] Md. S. I. Sumon et al., “CardioTabNet: a novel hybrid transformer model for heart disease prediction using tabular medical data,” Health Inf Sci Syst, vol. 13, no. 1, p. 44, Jul. 2025, doi: 10.1007/s13755-025-00361-7.
[16] Q. A. Hidayaturrohman and E. Hanada, “A Comparative Analysis of Hyper-Parameter Optimization Methods for Predicting Heart Failure Outcomes,” Applied Sciences, vol. 15, no. 6, p. 3393, Mar. 2025, doi: 10.3390/app15063393.
[17] Y. Rimal and N. Sharma, “Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy,” Multimed Tools Appl, vol. 83, no. 18, pp. 55091–55107, Nov. 2023, doi: 10.1007/s11042-023-17273-x.
[18] A. Banerjee et al., “Machine learning for subtype definition and risk prediction in heart failure, acute coronary syndromes and atrial fibrillation: systematic review of validity and clinical utility,” BMC Med, vol. 19, no. 1, p. 85, Dec. 2021, doi: 10.1186/s12916-021-01940-7.
[19] M. W. Segar et al., “Development and Validation of Machine Learning–Based Race-Specific Models to Predict 10-Year Risk of Heart Failure: A Multicohort Analysis,” Circulation, vol. 143, no. 24, pp. 2370–2383, Jun. 2021, doi: 10.1161/CIRCULATIONAHA.120.053134.
[20] M. Saqib et al., “Machine learning in heart failure diagnosis, prediction, and prognosis: review,” Annals of Medicine & Surgery, vol. 86, no. 6, pp. 3615–3623, Jun. 2024, doi: 10.1097/MS9.0000000000002138.