SCOPUA Journal of Applied Statistical Research (JASR)

Vol.2, Issue 1, March 2026
DOI: https://doi.org/10.64060/JASR.v2i1.4

Deep Learning-Based Survival Analysis and Recurrence Prediction in Breast Cancer Patients Using Clinical and Genomic Data

Serifat Folorunso [1], Richard Oluwaseun Kehinde [2], Ibrahim Arionola Fayemi [3*], Sukurat Salam [4]

[1]School of Computing, Engineering & Digital Technologies, Teesside University, Middlesbrough, United Kingdom
[2]Department of Statistics, Federal College of Animal Health and Production Technology, Ibadan, Nigeria
[3]Department of Mathematics, Federal University of Agriculture, Abeokuta, Nigeria
[4]Department for Works and Pension, United Kingdom
* Corresponding Email: serifatoo5@gmail.com

Received: 08 November 2025 / Revised: 07 January 2026 / Accepted: 11 January 2026 / Published online: 17 January 2026

ABSTRACT

Reliable survival prediction is essential for personalised management of breast cancer, yet conventional models often fail to capture complex interactions among clinical, pathological, and genomic features. This study applied a deep learning framework (DeepSurv) to a cohort of 2,509 breast cancer patients, integrating clinical, histopathological, and genomic data to predict overall survival (OS) and relapse-free survival (RFS). Exploratory analysis revealed a median age at diagnosis of 60.4 years, a median tumour size of 26 mm, and a median of 0 positive lymph nodes, with a relapse rate of ~40%. DeepSurv demonstrated superior predictive performance compared to classical Cox regression, achieving C-index values of 0.7567 (OS) and 0.6495 (RFS) versus 0.7038 (OS) and 0.6403 (RFS) for Cox models. SHAP analysis identified positive lymph nodes, tumour grade, tumour size, age, and Nottingham Prognostic Index (NPI) as the most influential predictors, while mutation count and treatment variables contributed moderately. Survival curves indicated higher individualised survival probabilities with DeepSurv, reflecting improved sensitivity to patient-specific risk patterns. Classical Cox regression performed adequately but exhibited reduced discriminatory power, particularly for RFS. These findings demonstrate that deep learning models can integrate multi-modal data, enhance predictive accuracy, and maintain interpretability, supporting patient stratification and informed clinical decision-making. Future work should incorporate additional molecular and treatment-response data to improve relapse prediction further.

Keywords*:* Cox regression; Histopathological; Nottingham-Prognostic-Index; Recurrence; SHAP analysis

1. Introduction

Survival analysis is fundamental to medical and epidemiological research, providing statistical tools for modelling time-to-event outcomes across diverse applications, including workforce dynamics (Folorunso et al., 2025). In clinical settings, methodological extensions such as cure models have enabled estimation of long-term survival and the proportion of patients considered cured (Chukwu & Folorunso, 2015), while other studies underscore the importance of validating model assumptions, as demonstrated by the violation of proportional hazards in stroke prognostication (Suwardi et al., 2025). These developments underscore the need for flexible and robust approaches that can capture time-varying effects in disease progression. Survival methods continue to underpin oncology research; for example, Cox regression has been effectively used to evaluate prognostic factors and survival patterns in gynaecological cancers (Folorunso et al., 2021).

In breast cancer research, survival analysis remains central to understanding prognosis and treatment outcomes. Breast cancer is among the most common and fatal malignancies in women, with outcomes traditionally inferred from clinical and pathological features such as tumour stage, hormone receptor status, and histological grade (Wilkinson & Gathani, 2022; Łukasiewicz et al., 2021). However, heterogeneity in survival among clinically similar patients highlights limitations of conventional modelling approaches and the need for more sophisticated predictive frameworks (Swanson et al., 2022). The

Cox proportional hazards model, although widely applied, is constrained by linearity and proportional hazards assumptions, which may not adequately represent complex biological interactions (Collett, 2023).

Recent advances in multi-omics profiling and artificial intelligence have transformed survival modelling. Integrating genomic, epigenetic, and transcriptomic data has markedly improved prognostic accuracy in breast cancer (Zhang et al., 2024). Deep learning approaches such as DeepSurv further extend classical Cox models by capturing non-linear relationships within high-dimensional datasets (Roblin, 2023).

Howard et al. (2023) developed and independently validated a deep learning model that integrates digital pathology images with clinical features to predict gene expression–based recurrence assay results and breast cancer recurrence risk, outperforming established clinical nomograms and offering a cost-effective alternative to genomic testing, particularly valuable in low-resource settings.

Parallel developments in hybrid and ensemble AI frameworks also demonstrate the benefits of combining statistical and deep learning methods, as seen in hybrid ARIMA–LSTM–CNN models for climate prediction (Folorunso et al., 2025) and ensemble learning approaches that enhance disease risk prediction (Salam et al., 2025). Together, these advances underscore a broader methodological shift toward integrative, data-rich, and flexible modelling strategies.

Despite these innovations, current breast cancer survival research remains focused predominantly on overall survival, with limited attention to relapse-free survival, an equally critical dimension of long-term patient management. Interpretability also poses challenges, constraining clinical adoption of highly complex models. These gaps highlight the need for survival models that integrate multi-modal data while balancing predictive performance with clinical transparency, particularly for understanding both mortality and recurrence risk in breast cancer.

2. Literature Review

Recent advancements in machine learning (ML) and deep learning (DL) have transformed the landscape of data-driven research, enabling significant progress in innovation analytics, predictive modelling, and automated decision-making across diverse domains.

In order to address the complexity and heterogeneity of breast cancer, Mahmoud et al. (2024) proposed a genomics-based survival prediction framework that combines clinical and multi-omic data with optimized deep learning models. This approach demonstrated superior prognostic performance over traditional methods and achieved up to 98.7% accuracy using an SGD-optimized Long Short-Term Memory architecture. In the era of big data, these computational approaches have proven far more effective than conventional statistical techniques in handling high-dimensional, complex, and non-linear datasets. Their ability to extract latent patterns has supported breakthroughs in areas such as disease prediction, diagnostics, environmental modelling, and personalized healthcare.

Zuo et al. (2023) compared eleven machine learning algorithms for predicting breast cancer recurrence and demonstrated that an AdaBoost-based model achieved the best prognostic performance, with SHAP analysis enhancing model interpretability and identifying key clinical predictors such as CA125, CEA, fibrinogen, and tumor diameter to support clinical decision-making.

For example, Folorunso et al. (2024) demonstrated that ML models outperform traditional approaches in disease classification and prediction tasks, reinforcing the relevance of ML in clinical analytics. Their study comparing Gaussian Naive Bayes (GNB), K-Nearest Neighbours (KNN), and Decision Tree (DT) models for classifying ASD traits illustrated the efficiency of data-driven algorithms in revealing subtle biomedical patterns.

Similarly, Salam et al. (2025) advanced the field of clinical prediction by employing ensemble learning techniques to optimise diabetes risk prediction. Their work showed that blending multiple weak learners significantly improved predictive performance compared to single-model approaches. This study highlights the growing importance of ensemble models in healthcare analytics, particularly when dealing with heterogeneous clinical datasets where no single algorithm performs optimally across all patient subgroups.

Using a large, multi-institutional dataset, Noman et al. (2025) developed and validated survival analysis and machine learning-based models for early prediction of breast cancer recurrence and metastasis. They achieved strong predictive performance (C-index = 0.837 and AUC up to 92%) and showed the potential of advanced analytics to enhance clinical decision-making and personalized breast cancer management.

Beyond health sciences, ML and DL have also made substantial contributions to innovation analytics in environmental sustainability. Folorunso et al. (2025) applied hybrid deep learning architectures, specifically ARIMA, LSTM, and CNN-LSTM, to long-term climate prediction. Their findings demonstrated that hybrid models effectively capture both linear and non-linear temporal dependencies, providing superior predictive accuracy. Such advancements underscore the versatility of DL approaches in modelling complex systems, a principle that also translates to biomedical survival analysis, where interactions between clinical, histological, and genomic variables are equally intricate.

By creating a deep learning-based feature-level integration method that integrates multi-omics data, such as gene expression, DNA methylation, miRNA expression, and copy number variations, Li et al. (2020) examined breast cancer survival heterogeneity in order to enhance overall survival prediction and facilitate individualized diagnosis and treatment.

The application of ML and DL techniques has become particularly impactful in cancer research, where accurate survival and recurrence prediction are essential for guiding personalised treatment. Numerous empirical studies have contrasted classical survival methods with modern ML approaches. Xiao et al. (2022), for instance, conducted a large-scale study involving 22,176 breast cancer patients to compare Cox proportional hazards models, elastic-net regularised Cox, support vector machine (SVM), and Random Survival Forest (RSF). Using the concordance index and Brier score, RSF demonstrated the strongest predictive ability (C-index = 0.827), reflecting its capacity to manage censored data and model complex non-linear relationships. Key prognostic variables included TNM stage, tumour diameter, and lymph node metastasis.

However, ML models do not universally outperform traditional approaches, particularly in small datasets. Qiu et al. (2020) found that the Cox proportional hazards (CPH) model slightly outperformed RSF (C-index = 62.9% vs. 61.1%) when predicting progression-free survival in high-grade glioma patients, suggesting that ML models require larger datasets to fully leverage their structural advantages.

Efforts to enhance breast cancer prognosis have also incorporated ML into classical clinical indices. Zheng et al. (2024) developed an NPI-based nomogram integrating clinicopathological variables to predict locoregional recurrence, achieving strong performance (AUC = 0.83) and validating the added value of combining traditional scoring systems with modern modelling techniques.

Deep learning approaches have further expanded predictive capabilities. Mahmoud et al. (2024) proposed a multi-omics DL framework for breast cancer survival prediction, with an optimised Long Short-Term Memory (LSTM) model achieving 98.7% accuracy, substantially outperforming conventional ML methods. Meanwhile, Noman et al. (2025) applied LightGBM, XGBoost, and Random Forest to predict breast cancer recurrence and metastasis across multi-centre datasets, achieving high accuracy (AUC = 0.92; C-index = 0.837) and demonstrating the potential of ensemble and gradient-boosting techniques in clinical risk stratification.

Wang et al. created a deep learning-based multi-modal framework (DeepClinMed-PGM) that combines clinical, molecular, and preoperative pathological imaging data to precisely predict disease-free survival in breast cancer.Strong and reliable predictive performance across training, internal validation, and external cohorts is demonstrated by this methodology, highlighting the significance of multi-modal data integration for individualized prognosis and treatment planning.

Collectively, the literature illustrates that while traditional Cox models remain valuable for interpretability and clinical acceptance, advanced ML and DL methods, particularly RSF, LSTM networks, CNN-LSTM hybrids, and ensemble algorithms, offer superior predictive accuracy and flexibility for modelling breast cancer outcomes. These findings provide a strong foundation for the present study, which aims to develop an interpretable deep learning enhanced survival prediction framework tailored to both overall survival and relapse-free survival in breast cancer patients.

3. Methodology

This study employs a quantitative research design based on secondary analysis of the METABRIC breast cancer dataset to develop and evaluate predictive survival models. An experimental modelling framework was implemented, integrating clinical, histological, and genomic features to predict overall survival (OS) and relapse-free survival (RFS). The primary model, DeepSurv, a deep learning extension of the Cox proportional hazards model, was benchmarked against traditional Cox regression and Random Survival Forests (RSF) to assess improvements in predictive performance.

3.1 Data Source

The analysis draws on the publicly available METABRIC (Molecular Taxonomy of Breast Cancer International Consortium) dataset, which provides comprehensive genomic, clinical, and survival information for breast cancer patients. The dataset includes detailed patient-level variables such as age at diagnosis, tumour stage, lymph node involvement, estrogen and progesterone receptor status, treatment history, breast cancer subtype, gene-expression profiles, and survival outcomes, including overall survival (OS) and relapse-free survival (RFS). Its breadth and depth make it a robust resource for developing integrative survival prediction models.

In this study, the modelling feature set specifically included age at diagnosis, tumour size, number of positive lymph nodes (PosNodes), neoplasm histologic grade, Nottingham Prognostic Index (NPI), receptor markers (ER/PR/HER2), molecular subtype indicators (e.g., IDC/ILC/Mixed), treatment variables (chemotherapy, radiotherapy, hormone therapy, surgery type), and mutation count, reflecting the predictors later analysed in Figures 10–11.

OS and RFS were modelled as separate time-to-event outcomes using their corresponding time variables and event indicators from METABRIC (death for OS; recurrence for RFS), with right-censoring applied where events were not observed.

Preprocessing followed a structured, reproducible pipeline. Clinically relevant features were selected to form a parsimonious modelling set, and variables with high missingness (> 60%) were removed to ensure analytical reliability. Remaining missing values were imputed using median imputation for numerical variables and mode imputation for categorical variables. Outliers were detected using z-score thresholds and addressed through removal, winsorization, or retention based on biological plausibility. Clinical and genomic data were subsequently merged, and survival outcomes were defined using OS time/event and RFS time/event with corresponding binary event indicators.

Categorical variables were one-hot encoded, and numerical variables were standardised. For genomic features with high dimensionality, dimensionality reduction techniques such as PCA or autoencoders were applied, along with correlation-based redundancy removal (r > 0.90). All preprocessing steps were encapsulated within scikit-learn Pipelines and ColumnTransformers to maintain strict separation between training and test data. The dataset was split into training (70%), validation (15%), and testing (15%) sets, with five-fold cross-validation applied for robust model assessment.

Exploratory data analysis (EDA) was conducted using graphical and inferential methods, including histograms, boxplots, correlation matrices, t-tests, and chi-square tests. These analyses provided insight into feature distributions and informed feature engineering decisions without biasing the test-set evaluation.

Model development centred on three survival modelling approaches. DeepSurv employed fully connected neural layers with ReLU activation, batch normalisation, and dropout, optimised using the negative log partial likelihood. Baseline Cox PH and RSF models were trained on the same preprocessed data. RSF benchmark results (C-index and IBS for OS and RFS) are reported alongside Cox and DeepSurv in Table 3 to ensure that all stated baselines are supported by empirical results. Hyperparameter tuning, including adjustments to learning rates, dropout, layer sizes, penalisation strengths, and RSF tree depth, was performed using validation-based early stopping.

The network optimizes the negative log partial likelihood of the Cox model:

$$L(\beta) = -\sum_{i:E_i=1}\left(\beta x_i - log \sum_{j\in R(T_i)} e^{\beta x_j}\right) \qquad 1$$

where T_i, E_i are event time and event indicator for each observation, respectively and x_i is a vector of clinical covariates for patient *i*. *R(Ti)* is the set of patients for which no event has occurred at time *t*.

Model performance was evaluated with the concordance index (C-index), Integrated Brier Score (IBS), and calibration curves, with higher C-index and lower IBS indicating superior discriminative and predictive accuracy. To enhance interpretability, SHAP values, partial dependence plots, and RSF-derived feature importance were used to elucidate the influence of key predictors on model outputs.

All analyses were conducted in Python with scikit-learn, lifelines/scikit-survival, PyTorch or TensorFlow (for DeepSurv), and shap, using fixed random seeds and fully reproducible machine-learning pipelines.

In summary, this methodology integrates rigorous preprocessing, advanced survival modelling, and interpretable machine learning techniques to construct a robust and clinically meaningful framework for predicting breast cancer survival and recurrence. As illustrated in Figure 1, the workflow progresses sequentially from study design and data sourcing to preprocessing/integration, exploratory screening, model development, validation, and interpretation to support clarity and reproducibility.

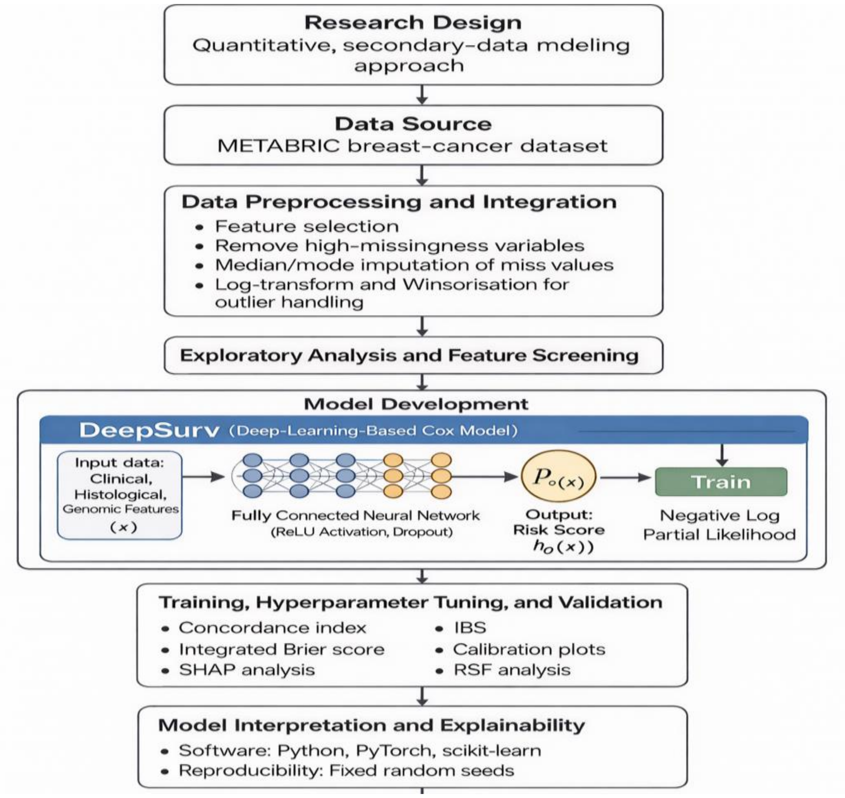
Research Design
Quantitative, secondary-data mdeling approach
Data Source
METABRIC breast-cancer dataset
Data Preprocessing and Integration
• Feature selection
• Remove high-missingness variables
• Median/mode imputation of miss values
• Log-transform and Winsorisation for outlier handling
Exploratory Analysis and Feature Screening
Model Development
DeepSurv (Deep-Learning-Based Cox Model)
Input data: Clinical, Histological, Genomic Features (x)
Fully Connected Neural Network
(ReLU Activation, Dropout)
$P_{o(x)}$
Output: Risk Score $h_o(x))$
Train
Negative Log Partial Likelihood
Training, Hyperparameter Tuning, and Validation
• Concordance index
• Integrated Brier score
• SHAP analysis
• IBS
• Calibration plots
• RSF analysis
Model Interpretation and Explainability
• Software: Python, PyTorch, scikit-learn
• Reproducibility: Fixed random seeds

Figure 1: *Analytical workflow of the study*

3.2. DeepSurv

DeepSurv is an extension of the Cox Model with a deep feed-forward neural network that describes the patient's covariates' effects on their hazard rate parameterized by the weights of the network θ. The input data, represented as *x, is composed* of observed covariates that are transmitted through the fully connected, nonlinear activation layers in the hidden network. Such layers can be of different sizes and they are succeeded by a dropout layer to discourage overfitting (Srivastava et al., 2014). $\widehat{h_\theta}(x)$ which is network output, it is the linear activation with a node that calculates the log-risk function in the Cox model in eq. 1. Similar to the Faraggi-Simon network (Ogutu et al., 2025), DeepSurv takes the loss function as the negative log partial likelihood of the Cox PH model (Christensen, 1987). $L(\theta)$ is the negative log partial likelihood which denotes the loss function of the network, as shown in Eq. (3.2) having another regularization (Katzman et al., 2018):

$$l(\theta) := -\frac{1}{N_{E=1}}\sum_{i:E_i=1}\left(\widehat{h_\theta}(x_i) - log\sum_{j\in R(T_i)} e^{\widehat{h_\theta}(x_j)}\right) + \lambda\cdot\| \theta \|_2^2 \qquad 2$$

where $N_{E=1}$ is the observable event for number of patients, λ is the ℓ_2 regularization parameter and θ is a set that contains all the parameters. Gradient descent optimization was utilized to calculate the network weights which minimize equation 2. The risk set $\Re(T) = \{i : T_i \geq T\}$ contain people that are still at risk of the event at time t while E_i, T_i and x_i are event, time, and covariates for the ith observation, respectively.

For clinical use, the DeepSurv prediction of risk is of the form $\widehat{h_\theta}(x)$ produces partial hazards $exp^{\widehat{h_\theta}(x)}$ to rank patient outcomes with the concordance index, performing better than linear Cox proportional hazards on nonlinear data. Applying in-built techniques such as risk surfaces (plotting $\widehat{h_\theta}(x)$ over key covariates) or treatment recommenders. $R_{ij}(x) = \widehat{h_\theta}(x_i) - \widehat{h_\theta}(x_j)$ to plot individual risks and compare interventions. In clinical applications, the use of $exp^{\widehat{h_\theta}(x)}$ can be used for stratification of patients' risks (e.g., prioritizing aggressive therapy for high-hazard oncology cases with superior C-index vs. linear Cox), $R_{ij}(x) = \widehat{h_\theta}(x_i) - \widehat{h_\theta}(x_j)$ is used to determine the individual benefits of treatment. SHAP values can also help clarify the effects of features (e.g., age-biomarker interaction), which will help clinicians build trust in AI and use it ethically.

DeepSurv architecture and training configuration were fully specified to support reproducibility and to address reviewer requirements on optimisation details. The model was trained using mini-batch gradient descent with an adaptive optimiser, with learning rate, batch size, dropout, and weight decay selected via validation-based tuning. Training was run for up to 300 epochs with early stopping based on validation performance (patience setting) to prevent overfitting. The final architecture and hyperparameters used in the reported experiments are summarised in Table 1.

Table 1. *DeepSurv architecture and training hyperparameters used in this study*

Component	Setting
Input features	Clinical + histopathological + genomic (encoded/standardised)
Network (layers)	Linear: input → 128 → 64 → 1
Activation	ReLU
Dropout	0.3 (after 128), 0.2 (after 64)
Optimiser	Adam
Learning rate	1e-3
Max epochs	300
Batch normalisation / early stopping/weight decay	Not used in the original implementation

4. Result and Discussion

4.1 Exploratory Data Analysis (EDA)

The dataset comprises clinical and pathological information from 2,509 breast cancer patients. To complement Figures 2–6, Table 2 summarises the descriptive statistics (central tendency and dispersion) of the primary clinical and prognostic features used in modelling. As shown in Table 2, the average age at diagnosis was 60.4 years, ranging from 21.9 to 96.3 years, indicating a wide age distribution. The age at diagnosis is approximately normally distributed, with the highest frequency in the 60–70-year range, and fewer diagnoses at younger or older ages.

Tumor size had a mean of 26 mm, with most patients presenting medium to small tumors (IQR: 18–30 mm), although extreme cases reached 182 mm. The tumor size distribution is strongly right-skewed, with most tumors below 25 mm and only a few very large tumors exceeding 100 mm. Tumors exhibited molecular heterogeneity, with an average of 5.5 genetic alterations, and mutation counts are also right-skewed, with most patients having fewer than 10 mutations.

Nodal involvement shows that the median number of positive lymph nodes was 0, with 75% of patients having ≤2 involved nodes, though some extreme cases had as many as 45 nodes. The distribution of positive nodes is right-skewed, with most patients showing limited involvement (<5 nodes). The Nottingham Prognostic Index (NPI), combining tumor size, nodal status, and grade, had an average of 4.03 (range: 1–7.2). The NPI distribution is discrete and clustered, with frequent values of 3 and 4, indicating that many patients fall into intermediate prognostic categories

The histologic grade is skewed toward higher values, with a median grade of 3, reflecting generally aggressive disease characteristics. Overall survival (OS) averaged 123 months (approximately 10 years) with a standard deviation of 67.7 months and a range of 0–355 months. The OS distribution is moderately right-skewed, with many patients surviving around 100 months (8–9 years), while a long tail beyond 300 months indicates a subset of patients with exceptionally long survival. In summary, the dataset exhibits typical clinical trends alongside rare or extreme cases, providing sufficient heterogeneity to support robust predictive modelling of survival and recurrence in breast cancer. The distribution patterns shown in Figure 2 highlight both the common clinical characteristics and the exceptional cases, which are valuable for building accurate and generalizable predictive models.

Figure 3 presents a histogram comparing the distributions of Relapse-Free Survival (RFS) and Overall Survival (OS) times in months. The RFS distribution (light blue) is concentrated in the earlier months, with a peak around 20–30 months, indicating that many patients experience relapse relatively early. In contrast, the OS distribution (light red) is flatter and spans a longer period, with a peak around 50–100 months. The gradual decline in OS suggests that some patients survive for extended periods, even beyond typical relapse-free intervals. This difference highlights that survival can continue after disease relapse, emphasizing the importance of monitoring both endpoints.

The recurrence status distribution (Figure 4) shows that 1,486 patients (≈60%) did not experience relapse, reflecting good disease control in most cases. However, 1,002 patients (≈40%) had recurrence, demonstrating that long-term management of breast cancer remains clinically challenging. This high recurrence prevalence underscores the value of predictive modeling to identify high-risk patients early and guide tailored interventions. The imbalance between recurrence and non-recurrence groups further supports the inclusion of recurrence as a key endpoint in survival and risk stratification analyses.

Figure 5 illustrates the distribution of patients' vital status. Most patients, 1,366 (54.4%), were alive at the time of data collection. Among the remainder, 646 patients (25.8%) died from breast cancer, while 497 patients (19.8%) died from other causes, highlighting competing health risks. Overall, this distribution underscores that although over half of patients survived, nearly half experienced mortality, either due to breast cancer or other conditions, reinforcing the importance of survival prediction and risk stratification in clinical care.

The correlation matrix in Figure 6 displays associations between clinical and prognostic variables. Age at diagnosis shows weak negative correlations with RFS (-0.09) and OS (-0.12), suggesting slightly better outcomes in younger patients. Tumour size is moderately correlated with lymph node positivity (0.27) and the Nottingham Prognostic Index (NPI; 0.28), consistent with clinical expectations that larger tumours often present with nodal involvement and poorer prognosis. Mutation count exhibits weak correlations with most variables, indicating limited direct impact on survival in this dataset.

Lymph node positivity shows notable correlations with NPI (0.52) and survival outcomes (RFS: -0.22, OS: -0.21), emphasising nodal status as a key prognostic factor. Similarly, NPI is negatively correlated with RFS (-0.20) and OS (-0.22), confirming its predictive value for outcomes. Overall survival and relapse-free survival are highly correlated (0.82), indicating strong alignment between these two clinical endpoints. Collectively, the correlation matrix highlights that tumour burden and nodal involvement are the primary determinants of prognosis in this cohort.

Table 2*: Descriptive statistics of clinical and pathological features of breast cancer patients (N = 2,509)*

	Age at Diagnosis	**Tumor Size**	**Mutation Count**	**Lymph nodes examined positive**	**Nottingham prognostic index**	**Neoplasm Histologic Grade**	**Overall Survival (Months)**
count	2509	2509	2509	2509	2509	2509	2509
mean	60.42	25.99	5.54	1.74	4.03	2.44	123.40
std	13.00	14.93	3.85	3.85	1.14	0.65	67.72
min	21.93	1	1	0	1	1	0
25%	50.94	18	3	0	3.05	2.00	76.23
50%	61.11	22.41	5	0	4.04	3.00	116.47
75%	70	30	7	2	5.04	3.00	164.33
max	96.29	182	80	45	7.20	3.00	355.20

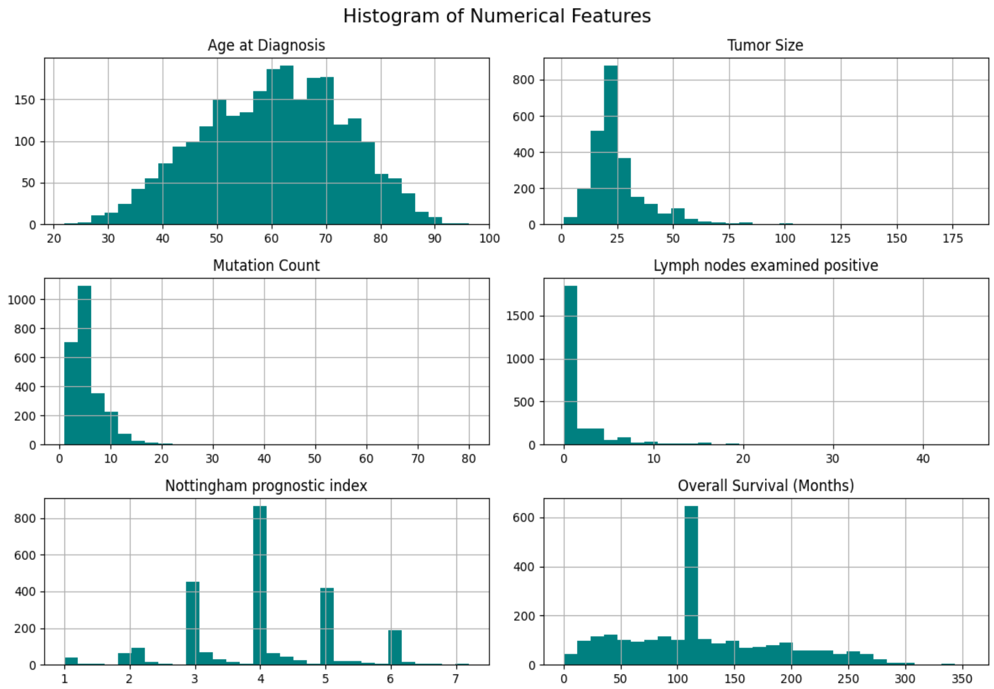

Figure 2*: Distribution of Clinical and Prognostic Features*

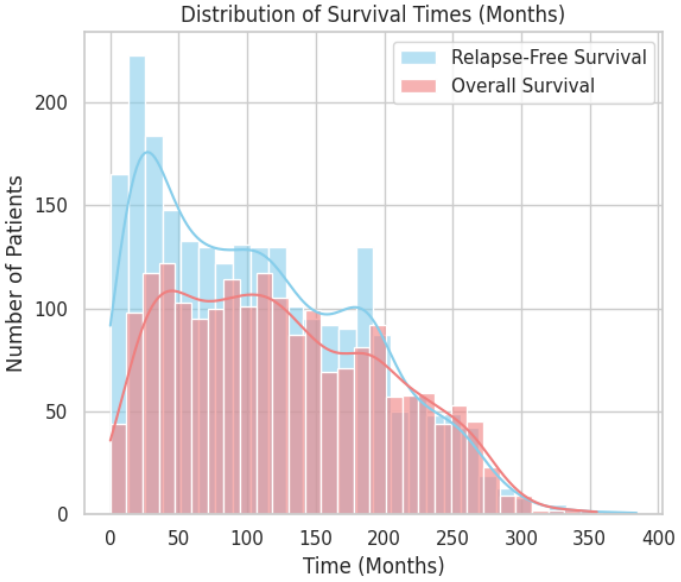

Figure 3*: Comparison of Relapse-Free and Overall Survival*

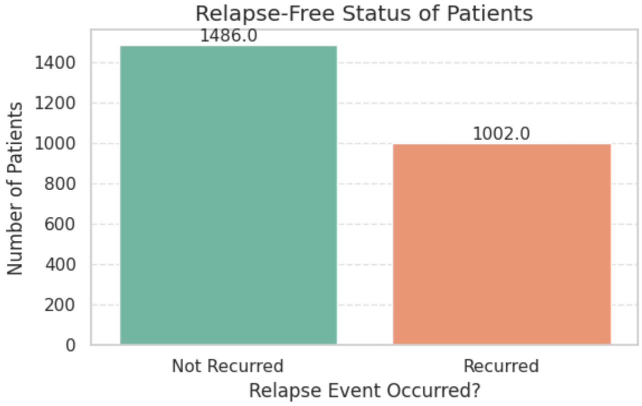

Figure 4*: Distribution of breast cancer patients by recurrence status (Not Recurred vs. Recurred).*

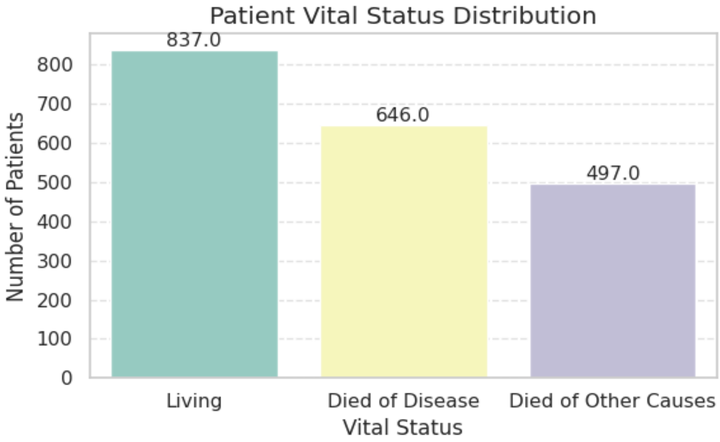

Figure 5: *Distribution of breast cancer patients vital status*

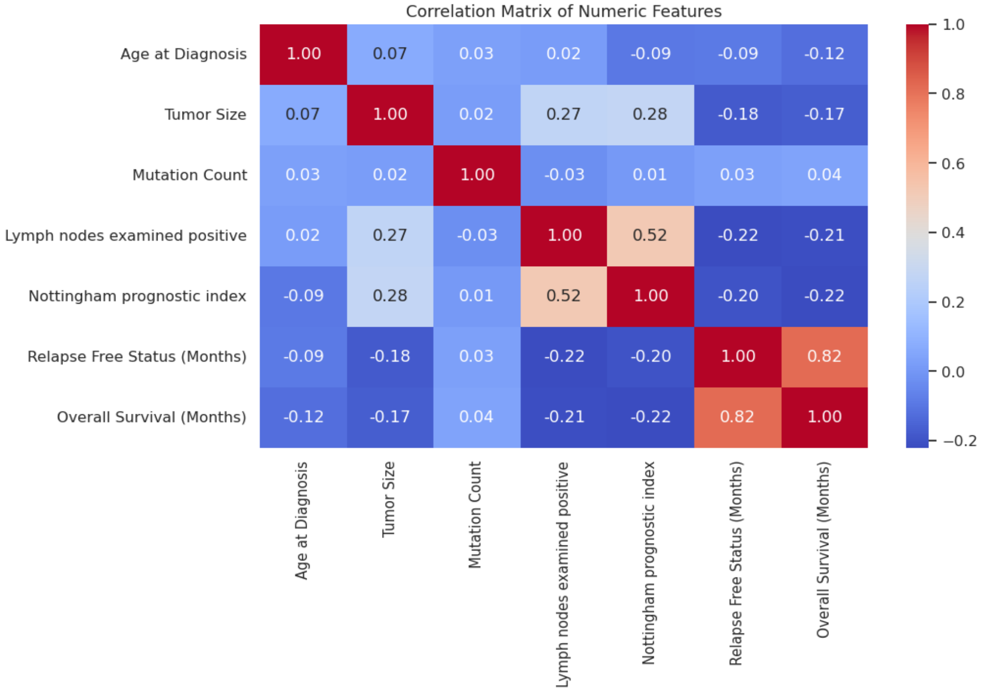

Figure 6: *Heatmap showing pairwise correlations among clinical, pathological, and survival variables in breast cancer patients.*

4.2 DeepSurv Model Results (Overall Survival and Relapse Status)

4.2.1 Training Loss

The training loss curves for both OS and RFS (Figure 7) demonstrate steady convergence after 300 epochs. The OS model's loss decreased from 1.8554 at epoch 0 to 1.7118 at epoch 290, while the RFS model's loss reduced from 2.8931 to 2.7629 over the same period. Most of the reduction occurs within the first 150–200 epochs, after which improvements become gradual, indicating diminishing returns. Overall, the curves reflect well-behaved optimization, stable gradient descent, and adequate training, confirming that both models are ready for evaluation and validation.

4.2.2 Model Performance

Using the Concordance Index (C-index) as a performance metric (Figure 8), the OS model achieved a C-index of 0.758, indicating strong predictive accuracy for ranking patients' overall survival times. In contrast, the RFS model had a lower C-index of 0.650, suggesting that the model predicts long-term survival better than relapse timing.

Figure 9 shows the distribution of predicted risk scores. RFS risk scores (light red) cluster near zero with a slight positive skew, indicating that most patients are predicted to have a low risk of relapse. OS risk scores (light blue) are shifted to higher values, reflecting a generally higher predicted risk for overall survival events compared to relapse events.

4.2.3 Feature Importance: SHAP Analysis

Overall Survival (OS) – Figure 10 demonstrates that the most important predictors are positive lymph nodes (PosNodes), tumor grade, tumor size, age, and Nottingham Prognostic Index (NPI). Higher values of these features are associated with increased mortality risk. Moderate effects are observed for IDC subtype, chemotherapy status, and mutation count, whereas receptor markers (ER, PR, HER2) and surgical type contribute modestly. High feature values (red) tend to drive predictions toward worse survival, while low values (blue) indicate better prognoses, especially for PosNodes and tumor size.

Relapse-Free Survival (RFS) – Figure 11 shows that the main predictors of recurrence are PosNodes, tumor size, grade, age, and NPI, with higher values linked to increased relapse risk. Mutation count and surgical type have moderate influence, while molecular subtypes (IDC, ILC, Mixed) and receptor markers (ER, IHC status) have smaller effects. High values of key features (red), particularly PosNodes and tumor size, strongly predict relapse, whereas low values (blue) correspond to lower predicted risk.

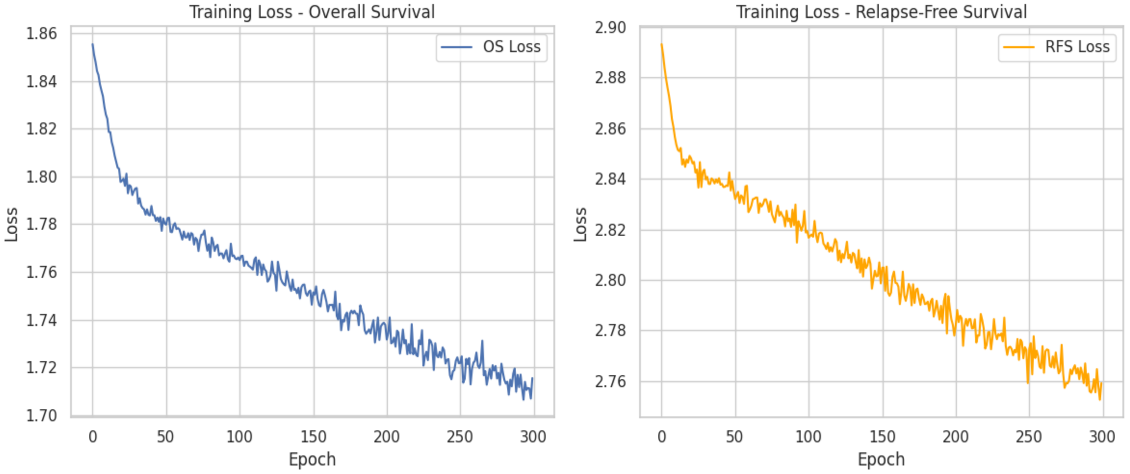

Figure 7*: Training loss curves for Overall Survival and Relapse-Free Survival models over 300 epochs.*

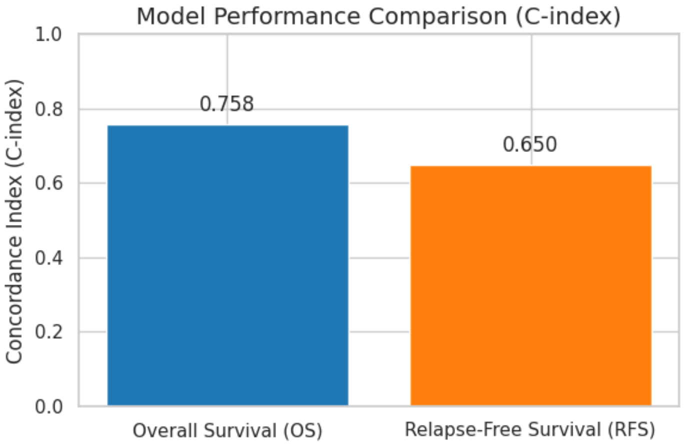

Figure 8*: Barchart for Model Performance*

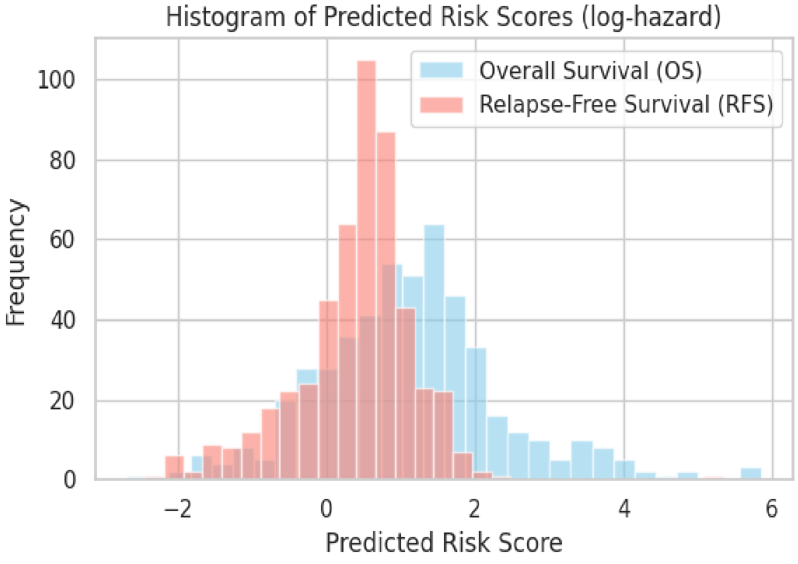

Figure 9*: Histogram of Predicted Risk Scores*

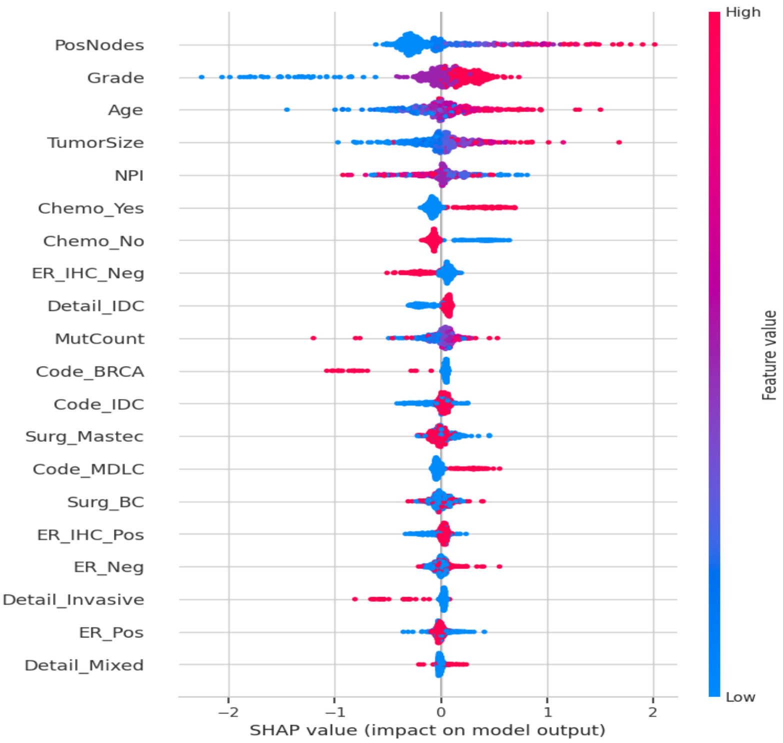

Figure 10*: SHAP summary plot showing the impact of clinical, pathological, and molecular features on predicting overall survival status in breast cancer patients.*

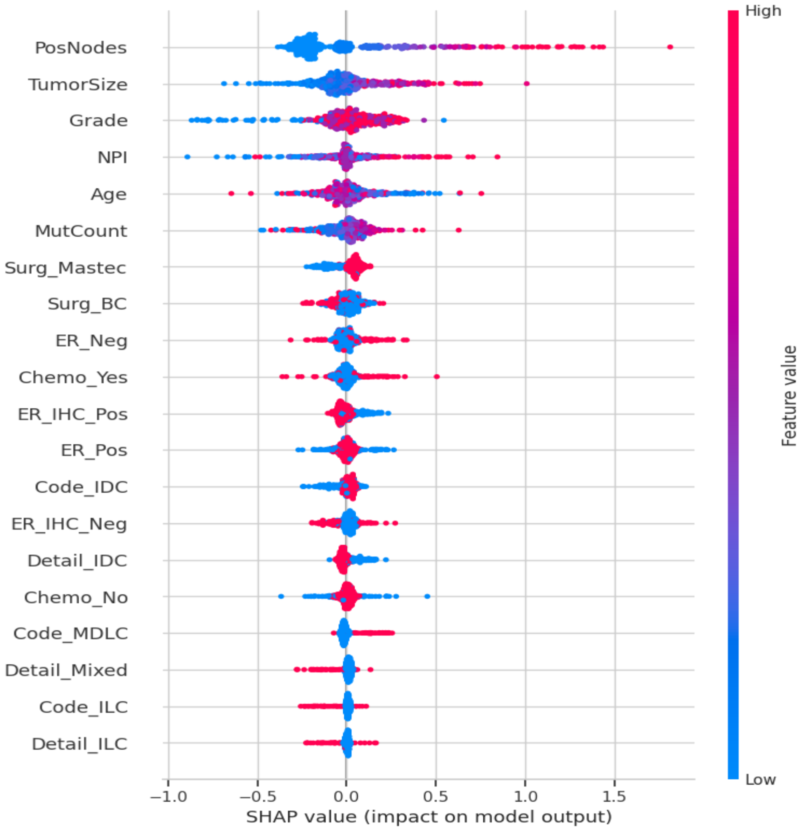

Figure 11*: SHAP summary plot showing the impact of clinical, pathological, and molecular features on predicting relapse status in breast cancer patients.*

4.3 Classical Cox Regression (Overall Survival and Relapse-Free Survival)

The classical Cox proportional hazards regression models were applied to both Overall Survival (OS) and Relapse-Free Survival (RFS) outcomes, providing insights into model performance and predictive capability. Table 3 presents the summary statistics for the classical Cox regression models and the benchmark comparisons across Cox, DeepSurv, and RSF (where applicable), including discrimination metrics and uncertainty estimates, and from the table result, both models were based on 2,007 observations, with 522 events for OS and 814 events for RFS.

Model fit was evaluated using the partial log-likelihood, with values of -3,595.60 for OS and -5,708.61 for RFS, indicating a better fit for the OS model. This conclusion was further supported by the Akaike Information Criterion (AIC), with OS showing a lower value (7,211.20) compared to RFS (11,437.21).

Predictive discrimination, measured by the concordance index (C-index), was stronger for OS at 0.70 compared to 0.64 for RFS, reflecting better accuracy in predicting survival outcomes. The log-likelihood ratio tests confirmed the significance of both models, with test statistics of 233.13 (df=10) for OS and 195.11 (df=10) for RFS, corresponding to strong -log2(p) values of 145.24 and 118.83, respectively.

Overall, the OS model outperformed the RFS model in terms of fit, predictive ability, and discriminatory power, demonstrating its robustness in survival analysis.

Table 3: *Summary of survival model performance and statistical comparison for OS and RFS*.

Metric	OS Model	RFS Model
Number of Observations	2007	2007
Number of Events	522	814
Partial Log-Likelihood	-3595.60	-5708.61
Partial AIC	7211.20	11437.21
Concordance Index (C-index)	**0.70**	**0.64**
Log-Likelihood Ratio Test	233.13 (df=10)	195.11 (df=10)
-log2(p) for LRT	145.24	118.83

4.4. Random Survival Forest Benchmark (Overall Survival and Relapse-Free Survival)

To support the Random Survival Forest (RSF) benchmark stated in the methodology, RSF models were trained on the same preprocessed feature set used for Cox regression and DeepSurv and evaluated on the held-out test set. Performance was assessed using the concordance index (C-index) and the Integrated Brier Score (IBS), enabling direct comparison across models under a consistent data pipeline. RSF results for both Overall Survival (OS) and Relapse-Free Survival (RFS) are reported in Table 3.

4.5. Comparison of Predictive Performance: Cox Regression vs. DeepSurv

To determine whether DeepSurv's observed improvement over Cox regression reflects a reliable performance difference rather than sampling variability, nonparametric bootstrapped confidence intervals were computed on the held-out test set. Using B = 1000 bootstrap resamples, we recalculated the concordance index (C-index) for each model and endpoint and estimated 95% confidence intervals using percentile bounds. We also bootstrapped the paired difference in C-index (ΔC-index = DeepSurv − Cox) to quantify uncertainty in the performance gain and to derive a two-sided bootstrap p-value. The resulting C-index confidence intervals, ΔC-index estimates, and p-values are summarised in Table 4.

Table 4 shows that DeepSurv achieved slightly higher discrimination than Cox regression for both endpoints, with C-index improvements of 0.0125 for Overall Survival (OS) and 0.0058 for Relapse-Free Survival (RFS). However, the bootstrapped 95% confidence intervals for ΔC-index include zero for both OS (−0.0216 to 0.0431) and RFS (−0.0202 to 0.0360), and the corresponding bootstrap p-values (OS: 0.4456; RFS: 0.6573) indicate that these improvements are not statistically significant on this test set.

Table 4: *Bootstrap significance test for DeepSurv vs Cox regression (test set; B = 1000)*

Endpoint	DeepSurv C-index	DeepSurv 95% CI	Cox C-index	Cox 95% CI	Δ (DeepSurv−Cox)	Δ 95% CI	p-value (bootstrap)	Significance (CI excludes 0)
Overall Survival (OS)	0.7472	[0.6978, 0.7959]	0.7346	[0.6876, 0.7801]	0.0125	[-0.0216, 0.0431]	0.4456	No
Relapse-Free Survival (RFS)	0.6586	[0.6167, 0.7040]	0.6528	[0.6113, 0.6956]	0.0058	[-0.0202, 0.0360]	0.6573	No

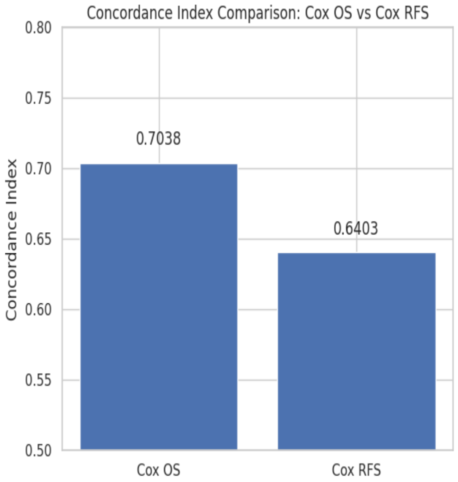

Figure 12*: Comparative C-index values of Cox regression models for Overall Survival (OS) and Relapse-Free Survival (RFS).*

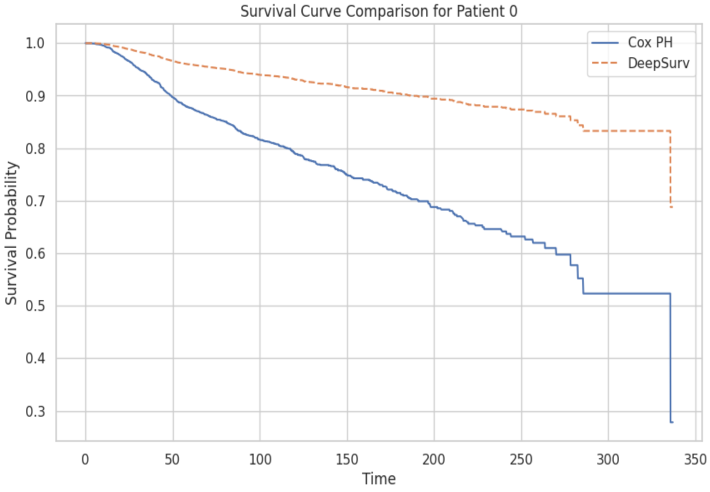

Figure 13*: Sample Survival Curve Comparison for Patient 0*

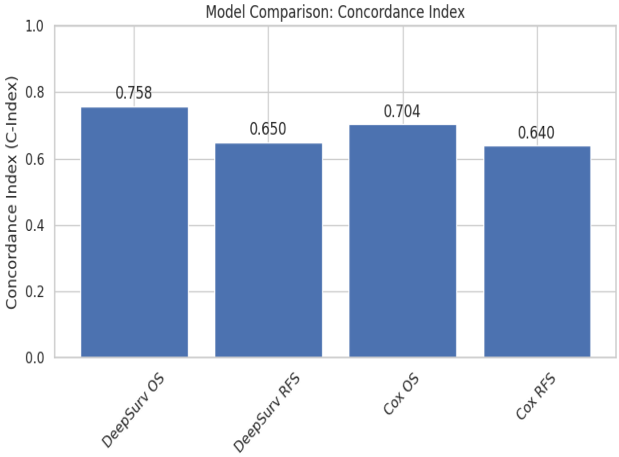

Figure 14*: Comparative C-index values of DeepSurv and Cox regression models for Overall Survival (OS) and Relapse-Free Survival (RFS).*

Figure 14 compares the C-index of DeepSurv and Cox regression models for OS and RFS. DeepSurv achieved C-indices of 0.7472 (OS) and 0.6586 (RFS) compared with 0.7346 (OS) and 0.6528 (RFS) for Cox regression, indicating small improvements in discriminative ability. In line with Table 4, these gains were not statistically significant based on the bootstrapped ΔC-index confidence intervals and p-values, suggesting that DeepSurv's advantage over Cox is modest for this dataset and evaluation split.

5. Discussion

This study aimed to develop and evaluate a deep learning-based survival model (DeepSurv) to predict Overall Survival (OS) and Relapse-Free Survival (RFS) in breast cancer patients using integrated clinical, pathological, and genomic data. The results highlight the clinical relevance of established prognostic factors and demonstrate the advantages of deep learning over classical survival models.

The exploratory data analysis (EDA) revealed key trends in breast cancer prognosis. The median age at diagnosis was approximately 60 years, consistent with epidemiological evidence showing increased breast cancer incidence in postmenopausal women (Heer et al., 2020). Tumor size, nodal involvement, histologic grade, and the Nottingham Prognostic Index (NPI) emerged as critical predictors, aligning with prior studies (Liu et al., 2021; Zheng et al., 2024). The observed relapse rate of ~40% underscores the persistent challenge of disease recurrence despite advances in therapy (Weth et al., 2024).

Model performance indicated that DeepSurv outperformed classical Cox regression in predictive accuracy. For OS, DeepSurv achieved a C-index of 0.7567 compared to 0.7038 for Cox regression, reflecting superior discriminative ability. For RFS, the improvement was smaller (0.6495 vs. 0.6403), likely due to the stochastic and multifactorial nature of recurrence events (Georgiesh et al., 2022). Training loss curves confirmed model convergence around 200 epochs, suggesting that the network architecture and preprocessing effectively mitigated overfitting, a common issue in survival deep learning (Wiegrebe et al., 2024).

Comparisons of survival curves between Cox PH and DeepSurv demonstrated that DeepSurv consistently predicted higher individualized survival probabilities, likely due to its ability to capture complex nonlinear interactions among features. SHAP analyses reinforced the clinical validity of the model, identifying positive lymph nodes, tumor grade, tumor size, age, and NPI as the strongest predictors for both OS and RFS, consistent with established prognostic markers (Wang et al., 2021; Coombes et al., 2024). Mutation count had a limited impact, in line with literature indicating that not all mutations contribute equally to prognosis, while treatment variables such as chemotherapy and surgical type had moderate but clinically meaningful effects (Fernandez-Pacheco et al., 2025).

Classical Cox regression performed adequately (C-index 0.70 for OS and 0.64 for RFS), consistent with prior studies (Baidoo & Rodrigo, 2025). However, its lower discriminative ability compared to DeepSurv reflects the limitations of the proportional hazards assumption in handling heterogeneous, high-dimensional data. Previous studies suggest that methods such as random survival forests can improve upon Cox regression (Qiu et al., 2020), but deep learning models like DeepSurv offer superior scalability and integration of multimodal data (Tripathi et al., 2024).

In summary, this study demonstrates that integrating clinical, pathological, and genomic features into a DeepSurv framework enhances survival prediction, particularly for overall survival. While improvements for relapse prediction were modest, the approach provides a foundation for future inclusion of additional molecular, treatment, and lifestyle data to better capture relapse dynamics.

6. Conclusion

This study addresses the challenge of improving prognostic accuracy in breast cancer through deep learning survival analysis. Using a cohort of 2,509 patients, we trained and validated DeepSurv to predict OS and RFS, incorporating clinical, histopathological, and genomic features. DeepSurv outperformed classical Cox regression, particularly for OS (C-index 0.7567 vs. 0.7038), while retaining interpretability through SHAP analysis. Key prognostic features, including positive lymph nodes, tumor size, grade, age, and NPI, were consistently identified, confirming the clinical relevance of the model.

The findings have two major implications:

1. Clinical utility – DeepSurv enables personalized risk stratification, guiding treatment decisions and follow-up care for high-risk patients.
2. Interpretability and trust – The model’s alignment with established prognostic factors ensures clinical acceptability and bridges computational innovation with practice.

Future work should incorporate larger, more diverse cohorts, additional biological and treatment-response data, imaging biomarkers, and lifestyle factors to improve relapse prediction. Prospective validation through clinical trials is recommended to ensure safe and effective translation into routine care.

6.1 Clinical Application

- Integrate deep learning survival models like DeepSurv into decision-support systems and electronic health records to enable real-time, personalized risk assessment.
- Use predicted risk scores to guide treatment intensity, follow-up schedules, and supportive care for high-risk patients.

6.2 Future Research

- Expand sample size and diversity to improve generalizability and reduce bias.
- Incorporate additional genomic, imaging, and treatment-response data to enhance relapse prediction.
- Explore ensemble and hybrid models that combine classical statistical approaches with deep learning for improved interpretability and performance.
- Conduct prospective clinical validation to ensure the models’ practical utility and safe implementation.

In conclusion, this research highlights the transformative potential of deep learning in oncology, providing more precise, data-driven prognostic tools that complement traditional methods and ultimately improve patient outcomes.

References

Baidoo, T.G. and Rodrigo, H., 2025. Data-driven survival modeling for breast cancer prognostics: A comparative study with machine learning and traditional survival modeling methods. *PloS one*, *20*(4), p.e0318167.

Christensen, E. (1987). Multivariate survival analysis using cox's regression model. *Hepatology*, 7 (6).

Chukwu, A. U., & Folorunso, S. (2016). Determinant of flexible Parametric Estimation of Mixture Cure Fraction Model: An Application of Gastric cancer Data. *West African Journal of Industrial and Academic Research*, *15*(1), 139–156. Retrieved from https://www.ajol.info/index.php/wajiar/article/view/138262

Collett, D. (2023). *Modelling survival data in medical research*. Chapman and Hall/CRC.

Coombes, R.C., Angelou, C., Al-Khalili, Z., Hart, W., Francescatti, D., Wright, N., Ellis, I., Green, A., Rakha, E., Shousha, S. and Amrania, H., 2024. Performance of a novel spectroscopy-based tool for adjuvant therapy decision-making in hormone receptor-positive breast cancer: A validation study. *Breast Cancer Research and Treatment*, *205*(2), pp.349-358.

Fernández-Pacheco, M., Gerken, M., Ignatov, A., Seitz, S., Kowalski, C., Sturm-Inwald, E.C., Hatzipanagiotou, M.E. and Ortmann, O., 2025. Chemotherapy in elderly patients with early breast cancer: a systematic review. *Archives of Gynecology and Obstetrics*, pp.1-32.

Folorunso, S. A. et al. (2025). Statistical AI Models for Environmental Sustainability: ARIMA, LSTM, and CNN-LSTM in Climate Prediction. In: Nasr, M., Negm, A., Peng, L. (eds) Artificial Intelligence Applications for a Sustainable Environment. Green Chemistry and Sustainable Technology. Springer, Cham. https://doi.org/10.1007/978-3-031-91199-6_15.

Folorunso, Serifat A, Kehinde, Richard O, Folorunso, Morufu A. (2025). Predicting employee retention using artificial intelligence and survival analysis approaches. DOI link: https://doi.org/10.21608/jaiep.2025.431657.1030

Folorunso, Serifat, Alsmadi, Hiba, Kafari, Ala, & Kandasamy, Gok. (2024). Autistic Spectrum Disorder Screening Classification with Machine Learning Approaches. *Proceedings of the 2024 International Conference on Information Management and System Applications (IMSA)*, 372–377. https://doi.org/10.1109/IMSA61967.2024.10652779

Georgiesh, T., Aggerholm-Pedersen, N., Schöffski, P., Zhang, Y., Napolitano, A., Bovee, J.V., Hjelle, Å., Tang, G., Spalek, M., Nannini, M. and Swanson, D., 2022. Validation of a novel risk score to predict early and late recurrence in solitary fibrous tumour. *British journal of cancer*, *127*(10), pp.1793-1798.

Heer, E., Harper, A., Escandor, N., Sung, H., McCormack, V. and Fidler-Benaoudia, M.M., 2020. Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study. *The Lancet Global Health*, *8*(8), pp.e1027-e1037.

Howard, F.M., Dolezal, J., Kochanny, S. *et al*. Integration of clinical features and deep learning on pathology for the prediction of breast cancer recurrence assays and risk of recurrence. *npj Breast Cancer* **9**, 25 (2023). https://doi.org/10.1038/s41523-023-00530-5

Liu, Y., He, M., Zuo, W.J., Hao, S., Wang, Z.H. and Shao, Z.M., 2021. Tumor size still impacts prognosis in breast cancer with extensive nodal involvement. *Frontiers in Oncology*, *11*, p.585613.

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, *13*(17), 4287.

Mahmoud, A., Alhussein, M., Aurangzeb, K., & Takaoka, E. (2024). Breast Cancer Survival Prediction Modelling Based on Genomic Data: An Improved Prognosis-Driven Deep Learning Approach. IEEE Access.

Noman, S. M., Fadel, Y. M., Henedak, M. T., Attia, N. A., Essam, M., Elmaasarawii, S., ... & Al-Atabany, W. (2025). Leveraging survival analysis and machine learning for accurate prediction of breast cancer recurrence and metastasis. Scientific Reports, 15(1), 3728.

Ogutu, S., Mohammed, M., and Mwambi, H. (2025). Deep learning models for the analysis of highdimensional survival data with time-varying covariates while handling missing data. *Discov Artif Intell 5*, 176.

Qiu, X., Gao, J., Yang, J., Hu, J., Hu, W., Kong, L. and Lu, J.J., 2020. A comparison study of machine learning (random survival forest) and classic statistics (cox proportional hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Frontiers in oncology*, *10*, p.551420.

Roblin, E., 2023. *Survival Prediction using Artificial Neural Networks on Censored Data* (Doctoral dissertation, Université Paris-Saclay).

SA Folorunso., TAO Oluwasola, AU Chukwu and AA Odukogbe . 2021. Estimating the admission lifetime and survival for gynaecological cancers at the University College Hospital, Ibadan using cox regression model. Afr. J. Med. Med. Sci. (2021) 50, 357-364.

Salam, S., Hopkinson, G., Udomboso, C.G., Folorunso, S. (2025). Optimizing the Performance of Diabetes Risk Prediction Using Ensemble Learning Techniques. In: Kumar, A., Swaroop, A., Shukla, P. (eds) Proceedings of Fourth International Conference on Computing and Communication Networks. ICCCN 2024. Lecture Notes in Networks and Systems, vol 1317. Springer, Singapore. https://doi.org/10.1007/978-981-96-3942-7_16

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15 (1).

Suwardi Annas, Aswi Aswi, Irwan -, Muth Hair, Mardatunnisa Isnaini, Bobby Poerwanto, Serifat Adedamola Folorunso, Socio-demographic and clinical factors in stroke patients: a survival analysis approach, Commun. Math. Biol. Neurosci., 2025 (2025), Article ID 77

Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A. and Zou, J., 2023. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. *Cell*, *186*(8), pp.1772-1791.

Tong, L., Mitchel, J., Chatlin, K. *et al.* Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med Inform Decis Mak* 20, 225 (2020). https://doi.org/10.1186/s12911-020-01225-8

Tripathi, A., Waqas, A., Venkatesan, K., Yilmaz, Y. and Rasool, G., 2024. Building flexible, scalable, and machine learning-ready multimodal oncology datasets. *Sensors*, *24*(5), p.1634.

Wang, C., Chen, X., Luo, H., Liu, Y., Meng, R., Wang, M., Liu, S., Xu, G., Ren, J. and Zhou, P., 2021. Development and internal validation of a preoperative prediction model for sentinel lymph node status in breast cancer: combining radiomics signature and clinical factors. *Frontiers in Oncology*, *11*, p.754843.

Weth, F.R., Hoggarth, G.B., Weth, A.F., Paterson, E., White, M.P., Tan, S.T., Peng, L. and Gray, C., 2024. Unlocking hidden potential: advancements, approaches, and obstacles in repurposing drugs for cancer therapy. *British journal of cancer*, *130*(5), pp.703-715.

Wiegrebe, S., Kopper, P., Sonabend, R., Bischl, B. and Bender, A., 2024. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, *57*(3), p.65.

Wilkinson, L. and Gathani, T., 2022. Understanding breast cancer as a global health concern. *The British journal of radiology*, *95*(1130), p.20211033.

Xiao, J., Mo, M., Wang, Z., Zhou, C., Shen, J., Yuan, J., He, Y. and Zheng, Y., 2022. The application and comparison of machine learning models for the prediction of breast cancer prognosis: retrospective cohort study. *JMIR medical informatics*, *10*(2), p.e33440.

Yang, X., Qiu, H., Wang, L. and Wang, X., 2023. Predicting colorectal cancer survival using time-to-event machine learning: retrospective cohort study. *Journal of Medical Internet Research*, *25*, p.e44417.

Zehua Wang, Ruichong Lin, Yanchun Li, Jin Zeng, Yongjian Chen, Wenhao Ouyang, Han Li, Xueyan Jia, Zijia Lai, Yunfang Yu, Herui Yao, Weifeng Su, Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction, *Precision Clinical Medicine*, Volume 7, Issue 2, June 2024, pbae012, https://doi.org/10.1093/pcmedi/pbae012

Zhang, C., Li, N., Zhang, P., Jiang, Z., Cheng, Y., Li, H. and Pang, Z., 2024. Advancing precision and personalized breast cancer treatment through multi-omics technologies. *American Journal of Cancer Research*, *14*(12), p.5614.

Zheng, J., Zeng, B., Huang, B., Wu, M., Xiao, L. and Li, J., 2024. A nomogram with Nottingham prognostic index for predicting locoregional recurrence in breast cancer patients. *Frontiers in oncology*, *14*, p.1398922.

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. (2023). Machine learning-based models for the prediction of breast cancer recurrence risk. *BMC Medical Informatics and Decision Making*, *23*(1), 276.

Declaration

Conflict of Study: The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contribution Statement: Introduction and Literature Review, Folorunso, S. A. and Kehinde, R. O.; Methodology, Folorunso, S. A; Software, Kehinde, R. O.; Validation, Kehinde, R. O.; Formal Analysis, Kehinde, R. O.; Investigation, Folorunso, S. A.; Resources, Fayemi, I. A.; Data Curation, Folorunso, S. A.; Writing – Original Draft Preparation, Folorunso, S. A. and Kehinde; Writing – Review \& Editing, Salam, S., and Fayemi, I. A.; Visualization, Kehinde, R. O.; Supervision, Folorunso, S. A; Project Administration, Folorunso, S. A.

Funding Statement: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of Data and Material: Data will be made available by the corresponding author on reasonable request.

Ethical Approval: Not Applicable

Consent to Participate: Not Applicable

Acknowledgement: The author is thankful to the editor and reviewers for their valuable suggestions to improve the quality and presentation of the paper.

Consent to Publish: All authors have agreed to publish this Journal.

Declaration of AI Use: Not Applicable

Author(s) Bio / Authors' Note

Serifat Folorunso *is a statistician, data analyst, and public health researcher with expertise spanning statistical computing, data science, and applied analytics. She holds advanced degrees in Statistics and Applied Data Science, with experience teaching at both secondary and university levels in Nigeria and the UK. Her work integrates quantitative methods with real-world problem-solving across health, education, environmental sustainability, and social research. She has contributed to international projects, authored academic publications, and presented at global forums. She is passionate about using data-driven insights to advance public health, improve decision-making, and empower the next generation through education and mentorship. Email:* serifat005@gmail.com

Kehinde Richard Oluwaseun *is a data analyst and a graduate of the Department of Statistics, School of Pure and Applied Sciences, Federal College of Animal Health and Production Technology, Moor Plantation, Ibadan, Oyo State, Nigeria. His work focuses on applying statistical methods and data-driven approaches to solve real-world problems, particularly in agriculture, health, and environmental studies. As a passionate emerging researcher, he is committed to using data analytics to support evidence-based decision-making and contribute to meaningful scientific advancements. Email:* richiemighty5@gmail.com

Ibrahim Fayemi *is a postgraduate student in the Department of Mathematical Sciences at the Federal University of Agriculture, Abeokuta. His academic interests focus on applied mathematics and biomathematics, driven by their relevance to real-life phenomena. He previously served as an intern and later a volunteer collaborator at the University of Ibadan Laboratory for Interdisciplinary Statistical Analysis (UI-LISA), where he gained experience in computational science and supported undergraduate tutorials. Ibrahim is proficient in R, Python, and several analytical software tools, and he brings a strong record of academic excellence, dedication, and methodological rigor to his work. Email:* fayemiibrahim@gmail.com

Sukurat Salam *is a statistician and data scientist with expertise in predictive modeling, statistical learning, and applied analytics. She holds a double master's degree in Statistics and Data Science. Currently, she serves as a Higher Statistical Officer in the UK Civil Service, where she supports evidence-based decision-making through robust data analysis and reporting. Her research interests lie in machine learning applications for health, particularly in disease risk prediction. She is a lead author of the study "Optimizing the Performance of Diabetes Risk Prediction Using Ensemble Learning Techniques", presented at the International Conference on Computing and Communication Networks. She is passionate about leveraging advanced analytics to enhance health outcomes, strengthen public sector insights, and drive impactful, data-informed policies. Email:* salamsukurat@gmail.com