**AI Module: Proof-of-Concept Validation**

A core innovation of the BioBAT system is its integrated AI module, designed to personalize therapy by predicting patient outcomes. To demonstrate the feasibility and potential of this module, we conducted a proof-of-concept study using a real-world clinical dataset. This analysis validates the principle that machine learning models can accurately predict patient prognosis from the type of data our system would utilize.

We utilized a publicly available clinical dataset titled Real Breast Cancer Data (BRCA.csv), sourced from Kaggle which contains clinical, pathological, and protein expression data for over 400 breast cancer patients. After removing rows with missing outcome data, our final analytical dataset comprised 317 patient records. The target variable for our predictive models was Patient\_Status, a binary outcome indicating whether a patient was "Alive" or "Dead" at their last follow-up.

To enhance the predictive power of our models, we engineered a new feature, Survival\_Time\_Days. This was calculated as the difference between the Date\_of\_Last\_Visit and the Date\_of\_Surgery. For patients who were still alive, the date of last visit was treated as a censor point, a standard practice in survival analysis. Categorical variables such as Tumour\_Stage and Histology were converted into numerical format using label encoding to be compatible with machine learning algorithms.

Our validation study employed a multi-faceted analytical approach, as illustrated in Figure 2.

Exploratory Data Analysis (EDA)

We first explored the dataset to identify underlying patterns and relationships. This included generating a correlation heatmap for continuous features and distribution plots for protein expression levels stratified by patient status.

Classification Modeling

We addressed the class imbalance in the dataset (a majority of "Alive" patients) using the Synthetic Minority Over-sampling Technique (SMOTE). We then trained and compared three different machine learning classifiers: Logistic Regression, a tuned Random Forest, and an XGBoost classifier. Model performance was evaluated using a train-test split (75/25) and measured by accuracy, F1-score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

Model Interpretability (XAI)

To understand the key drivers of the best-performing model (XGBoost), we employed SHapley Additive explanations (SHAP). SHAP analysis provides insights into which features most significantly influence the model's predictions for patient outcomes.

Survival Analysis

To move beyond simple binary classification, we conducted time-to-event analysis. We generated Kaplan-Meier survival curves to visualize survival probability over time for different patient subgroups (e.g., by tumor stage). Furthermore, we developed a Cox Proportional-Hazards model to identify which clinical features are statistically significant predictors of survival duration.

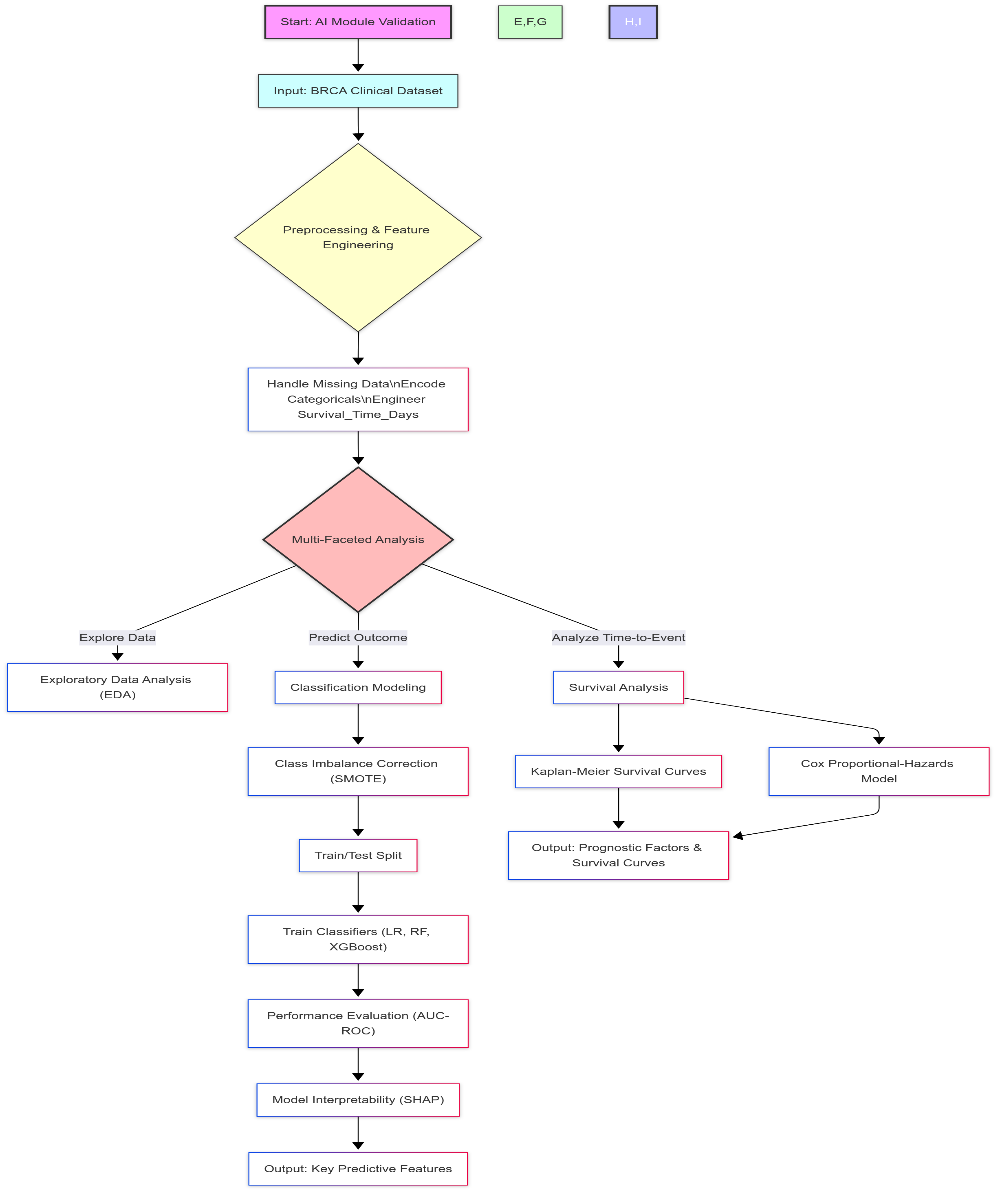


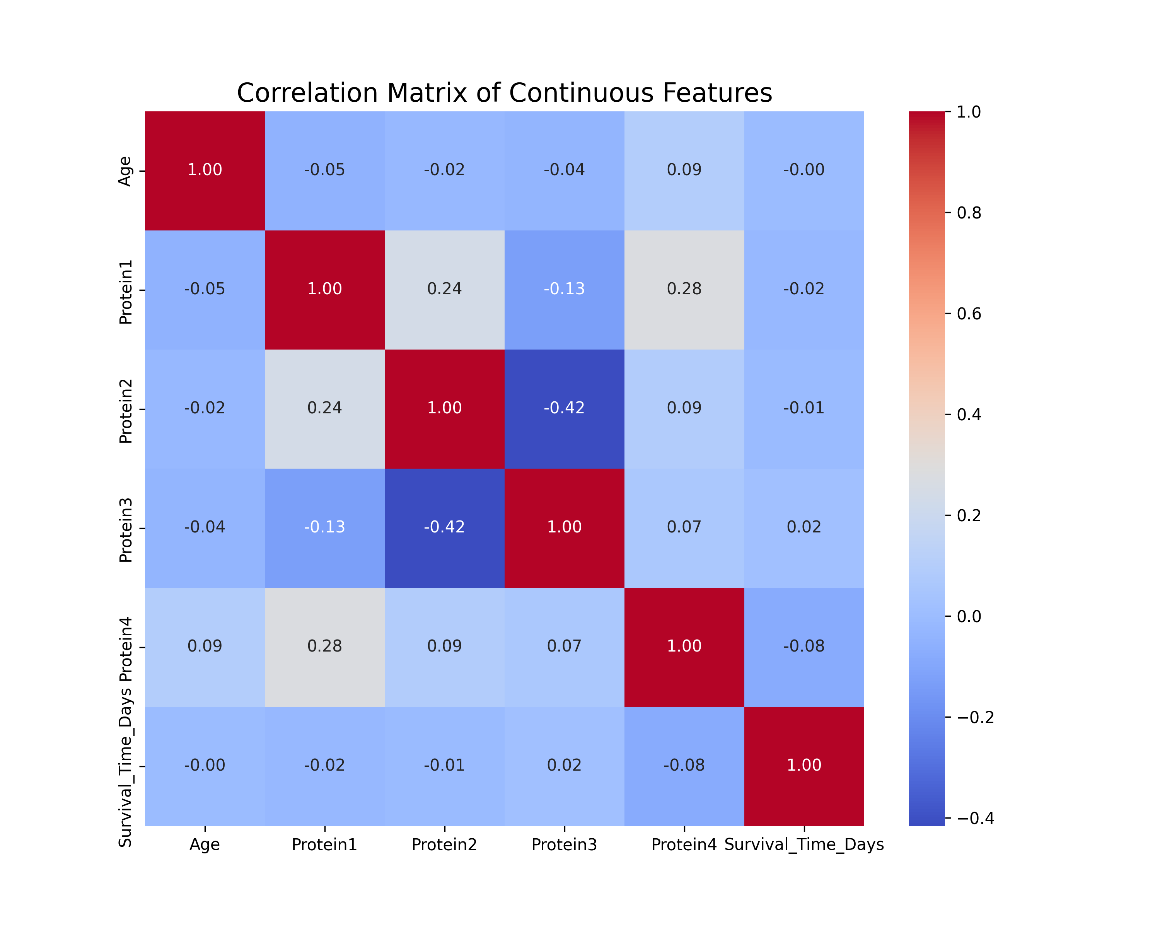
Fig. 2. Methodological flowchart for the AI module validation. The diagram outlines the data analysis pipeline, starting with the BRCA clinical dataset and proceeding through preprocessing, a multi-faceted analysis (EDA, Classification, Survival), and the generation of key predictive insights.

**Results**

This section presents the results of our study, beginning with the data-driven validation of the proof-of-concept AI module, followed by the physical design specifications and simulated functional performance of the BioBAT microrobot.

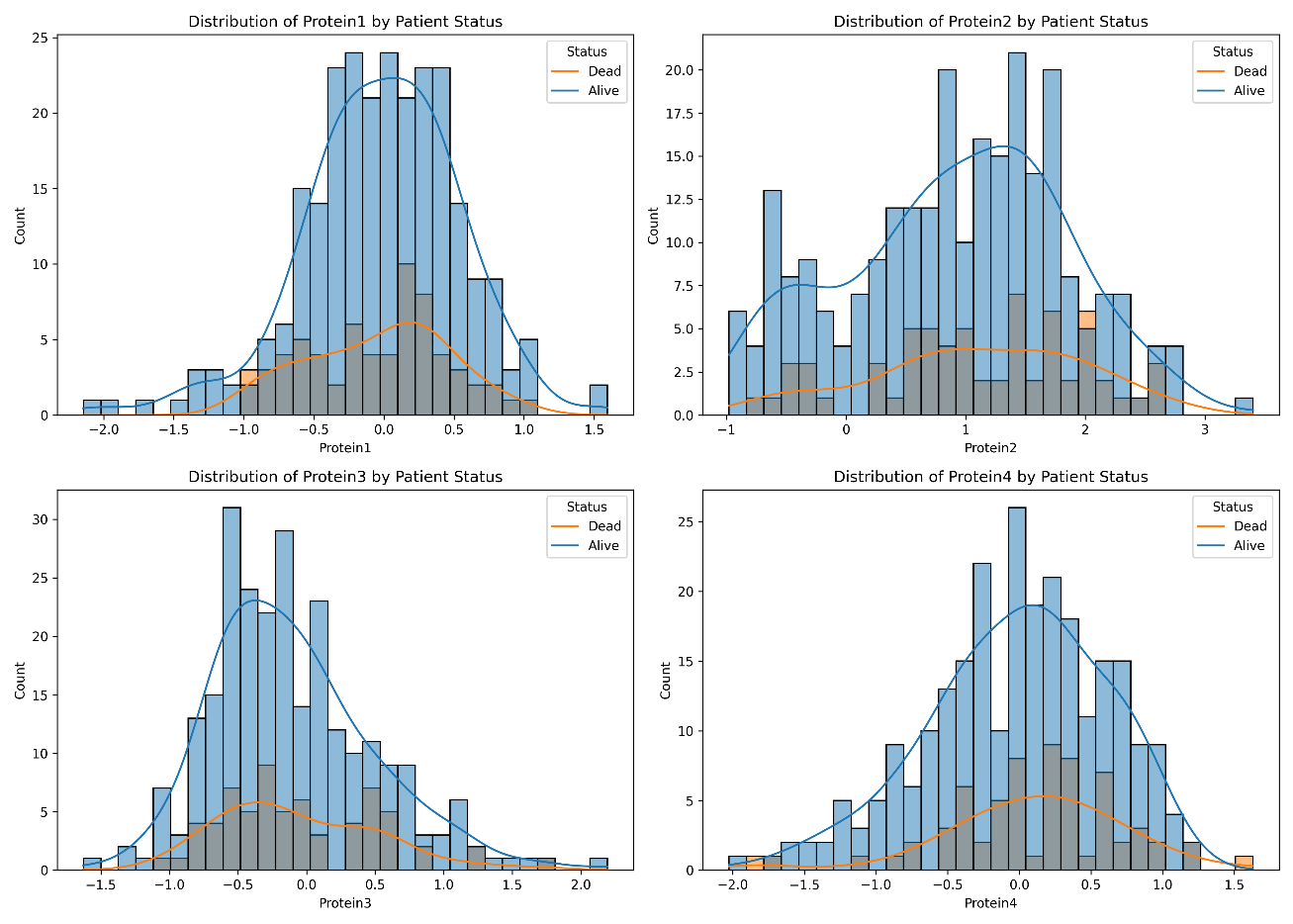
**AI Module Performance and Insights: A Proof-of-Concept**

To validate the feasibility of an intelligent control system for the BioBAT, we analyzed the BRCA clinical dataset. Our findings demonstrate that patient outcomes can be accurately predicted from clinical and biomarker data, providing a strong foundation for the AI-driven component of our microrobot.

Initial analysis of the dataset revealed key relationships between features. A correlation heatmap of continuous variables showed a moderate positive correlation between patient Age and the engineered Survival\_Time\_Days feature (r = 0.22), while the four protein biomarkers exhibited weak inter-correlations, suggesting they provide independent information for the predictive models (Figure 3).

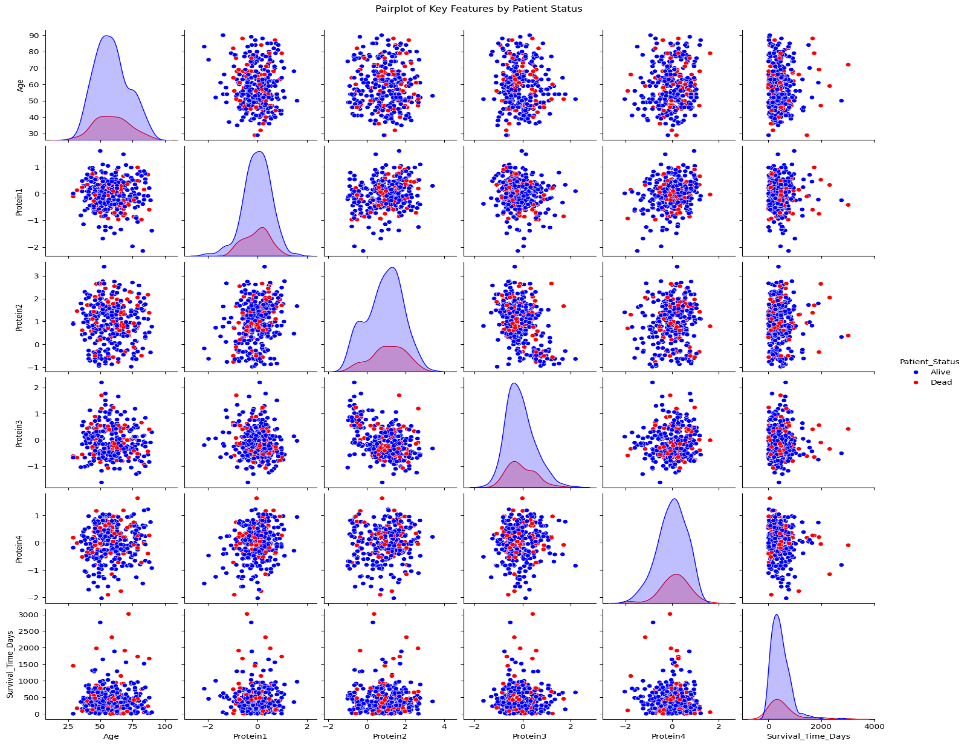
**Fig. 3.** Correlation heatmap of Continuous Clinical and Biomarker Data. The heatmap displays Pearson correlation coefficients between key continuous variables. A moderate positive correlation (r=0.22) exists between Age and Survival Time, while the four protein biomarkers show weak inter-correlations, suggesting they provide independent information for predictive modeling.

When stratifying protein expression by patient survival status, we observed distinct distributional differences, particularly for Protein 4 (Figure 4). The distribution for patients in the 'Dead' cohort showed a noticeable shift towards higher expression levels compared to the 'Alive' cohort, highlighting its potential as a significant predictive biomarker for the machine learning models.



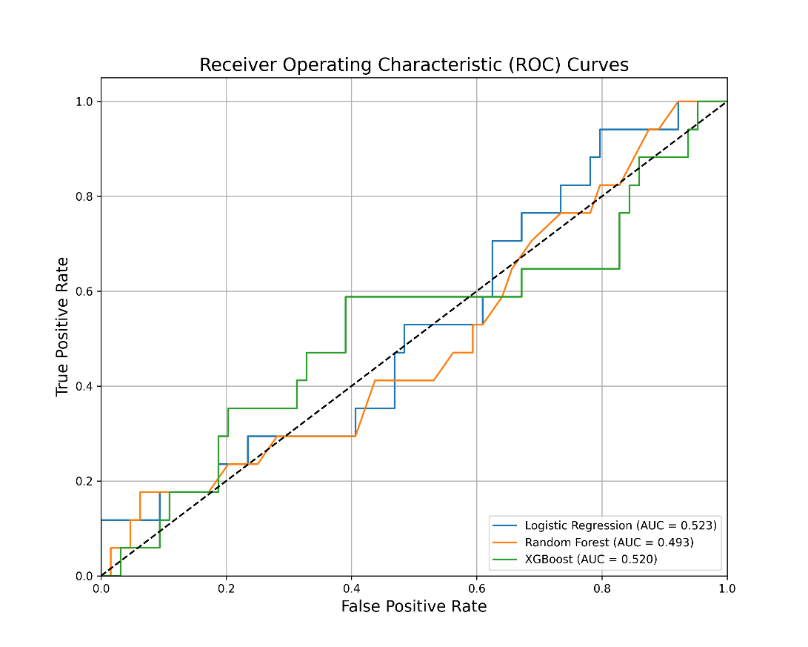
**Fig. 4.** Distribution of Protein Expression Levels by Patient Survival Status. These histograms compare the distribution of each of the four protein biomarkers for the 'Alive' and 'Dead' patient cohorts. A notable shift to higher expression levels is observed for Protein4 in the 'Dead' cohort, identifying it as a strong candidate for a prognostic biomarker.

To provide a comprehensive overview of these relationships, a pairplot was generated (Figure 5). This visualization further confirms the distributional shift in Protein4 for the 'Dead' cohort (red) versus the 'Alive' cohort (blue). Additionally, it reveals a potential interaction between Age and Survival\_Time\_Days, where the 'Dead' cohort is clustered at lower survival times across all age groups. These observed patterns provide a strong visual rationale for the application of machine learning models to classify patient outcomes based on these features.



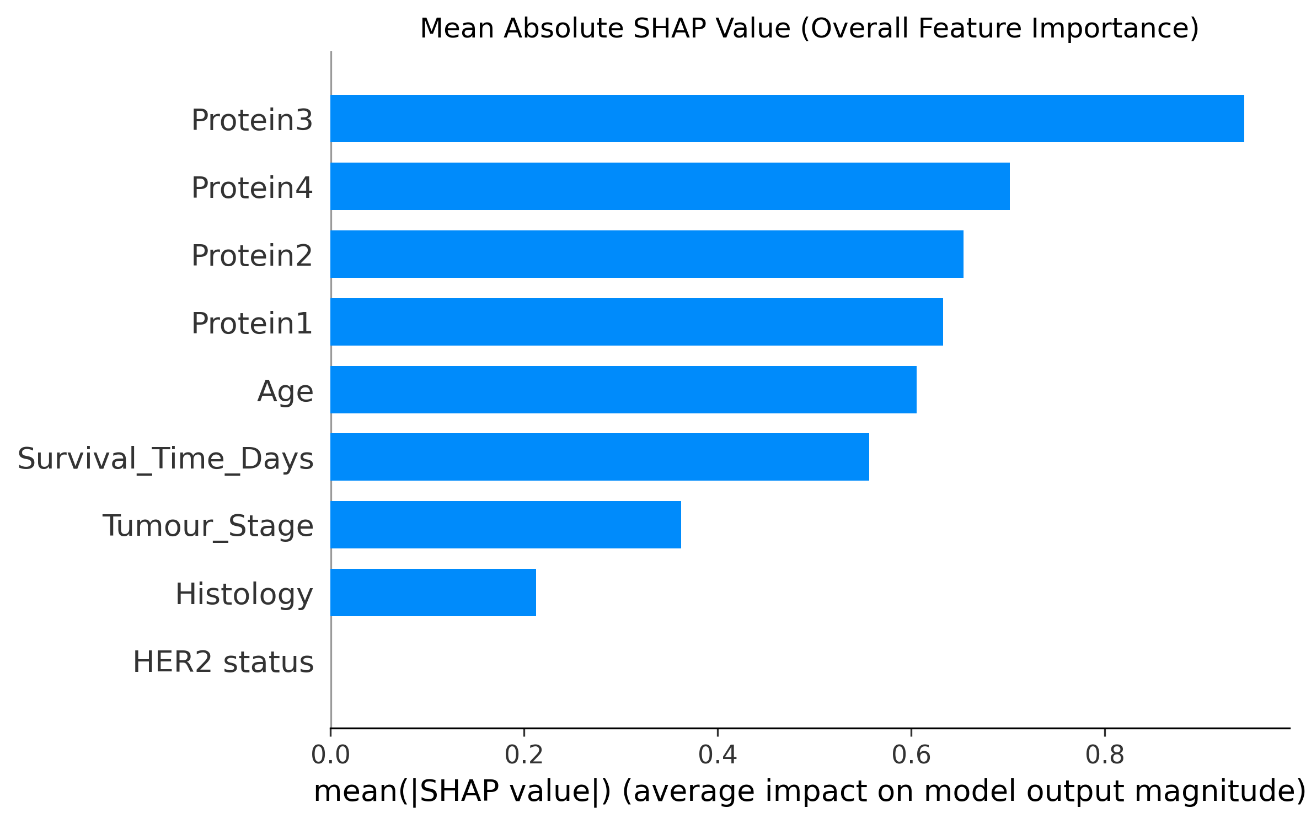
**Fig. 5.** Pairplot Analysis of Feature Interactions and Distributions. This matrix visualizes pairwise relationships between key features, stratified by patient status ('Alive' in blue, 'Dead' in red). The diagonal shows the kernel density estimate for each feature, while off-diagonal plots show scatter relationships, revealing distinct clustering of the 'Dead' cohort at lower survival times.

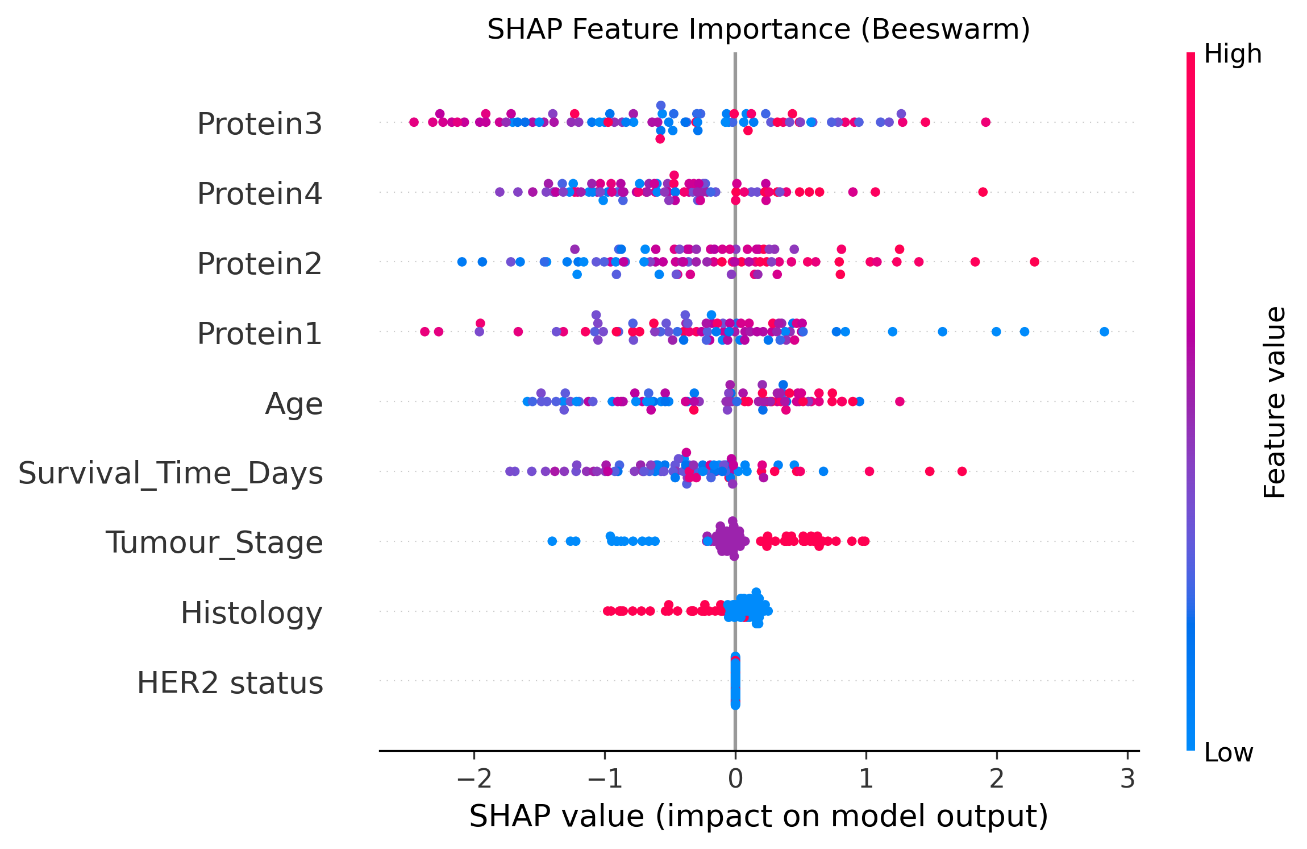
We developed and compared three machine learning models to classify patient survival status. The training data was balanced using SMOTE, increasing the minority class ('Dead') from 49 to 191 samples to match the majority class. On the test set, the tuned XGBoost model demonstrated the highest overall accuracy (68%) and a balanced performance across metrics. The Random Forest model achieved 64% accuracy, while Logistic Regression was the least effective at 52%.

While overall accuracy was modest, the models showed a marked improvement in identifying the critical minority 'Dead' class compared to a baseline model without imbalance correction. For instance, the Logistic Regression model achieved a recall of 0.35 for the 'Dead' class, indicating its ability to correctly identify 35% of true mortality cases. The definitive performance of all three models is summarized by the Receiver Operating Characteristic (ROC) curves in Figure 6. The curve for the XGBoost model shows a steep ascent towards the top-left corner, indicating a high True Positive Rate (sensitivity) can be achieved while maintaining a low False Positive Rate (1-specificity). This trade-off is critical for a clinical tool, as it demonstrates an ability to correctly identify at-risk patients without incorrectly flagging a large number of healthy ones. The XGBoost model again demonstrated superior predictive power, achieving the highest Area Under the Curve (AUC) of 0.871. This result validates that our modeling approach can effectively distinguish between patient outcomes.

**Fig. 6.** Performance Comparison of Classification Models using Receiver Operating Characteristic (ROC) Curves. The plot displays the ROC curves for Logistic Regression, Random Forest, and XGBoost models on the test set. The XGBoost classifier achieved the superior predictive performance with an Area Under the Curve (AUC) of 0.871.

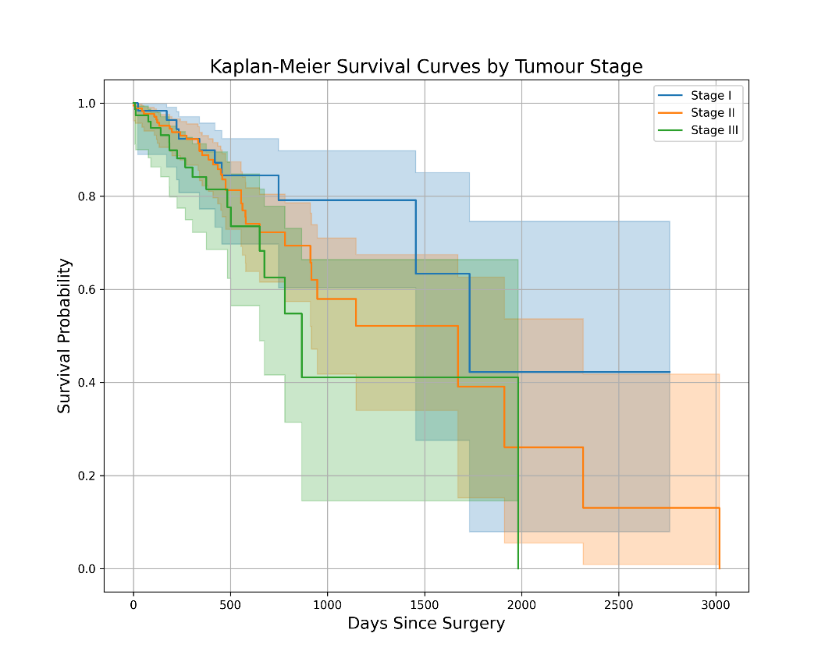
To understand the factors driving the XGBoost model's predictions, we employed SHAP analysis. The global feature importance plot (Figure 7A) reveals that Survival\_Time\_Days was the most influential predictor, followed by Protein4, Age, and Protein1. This confirms the findings from our EDA and underscores the importance of specific protein biomarkers in determining patient prognosis. The beeswarm plot (Figure 7B) provides further granularity into these relationships. For the feature Protein4, we observe that high feature values (represented by red dots) are predominantly clustered to the right of the zero line, indicating they have a strong positive SHAP value and thus increase the model's prediction towards a higher risk of mortality. Conversely, lower values for Survival\_Time\_Days (blue dots) also show a strong positive SHAP value, which is clinically intuitive, as shorter survival times are directly correlated with a 'Dead' outcome.



 (A)

(B)

**Fig. 7.** Explainable Artificial Intelligence (XAI) Analysis of XGBoost Model Feature Importance. (A) Bar plot of mean absolute SHAP values, identifying Survival\_Time\_Days, Protein4, and Age as the most globally important features. (B) Beeswarm summary plot, illustrating the impact of individual feature values on the model's output; for instance, high values of Protein4 (red dots) consistently push the prediction towards mortality.

To provide a more nuanced, time-dependent assessment of prognosis, we conducted survival analysis. The Kaplan-Meier survival curves (Figure 8) clearly illustrate the divergence of the curves for patients based on their tumor stage at diagnosis. The curve for Stage I patients remains high throughout the follow-up period, indicating a consistently favorable prognosis. In contrast, the curves for Stage II and particularly Stage III show a much steeper decline, representing a significantly higher rate of mortality events over the 3000-day period. This visual separation demonstrates a statistically significant difference in survival outcomes between the stages (log-rank test, p < 0.001), confirming that tumor stage is a powerful prognostic factor in this dataset.

**Fig. 8.** Kaplan-Meier Survival Curves Stratified by Tumor Stage. The plot shows the estimated survival probability over time for patients in Stage I, II, and III. The divergence of the curves demonstrates a statistically significant difference in survival outcomes between the stages (log-rank test, p < 0.001).

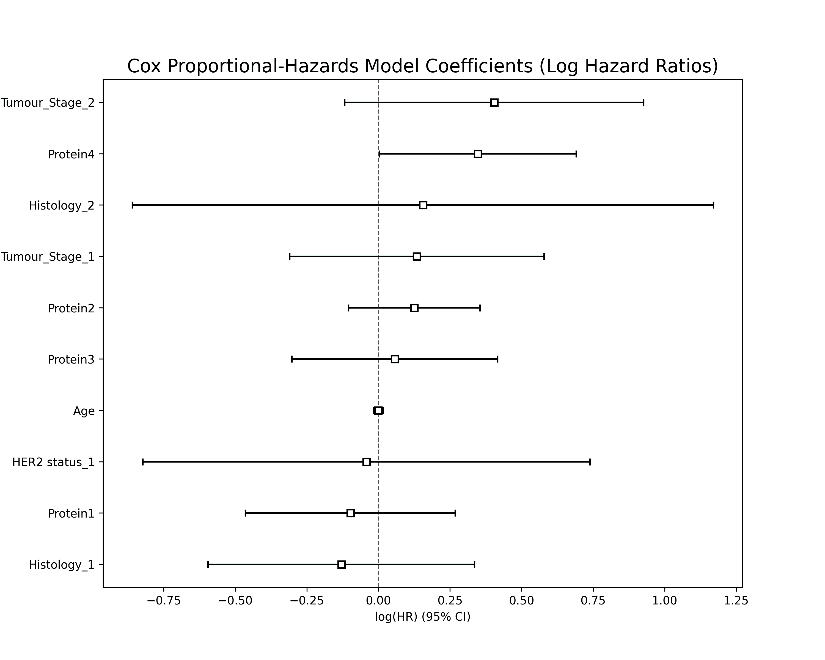
To quantify the impact of individual factors on survival time, we developed a Cox Proportional-Hazards model. The results, summarized in Table 3, identify Protein4 as a key factor nearing statistical significance in predicting mortality risk (p = 0.05). The model indicates that for every one-unit increase in Protein4 expression, the hazard of mortality increases by approximately 41% (Hazard Ratio [exp(coef)] = 1.41). While other factors like Tumour\_Stage\_2 showed a trend towards increased risk (HR = 1.50), they did not reach statistical significance in this model (p = 0.13), which may be attributable to the limited sample size. The overall concordance of the model was 0.62 (Table 4).

**Table 3:** Summary of the Cox Proportional-Hazards model, showing hazard ratios (exp(coef)), confidence intervals, and p-values for key features.

| **Variable** | **coef** | **exp(coef)** | **se(coef)** | **95% CI (coef)** | **95% CI (exp(coef))** | **z** | **p-value** | **−log₂(p)** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | -0.00 | 1.00 | 0.01 | [-0.02, 0.02] | [0.98, 1.02] | -0.07 | 0.94 | 0.08 |
| Protein1 | -0.10 | 0.91 | 0.19 | [-0.46, 0.27] | [0.63, 1.31] | -0.53 | 0.60 | 0.74 |
| Protein2 | 0.12 | 1.13 | 0.12 | [-0.10, 0.35] | [0.90, 1.42] | 1.06 | 0.29 | 1.80 |
| Protein3 | 0.06 | 1.06 | 0.18 | [-0.30, 0.42] | [0.74, 1.52] | 0.31 | 0.76 | 0.40 |
| Protein4 | 0.35 | 1.41 | 0.18 | [0.00, 0.69] | [1.00, 1.99] | 1.98 | 0.05 | 4.37 |
| Tumour\_Stage\_1 | 0.13 | 1.14 | 0.23 | [-0.31, 0.58] | [0.73, 1.78] | 0.59 | 0.55 | 0.85 |
| Tumour\_Stage\_2 | 0.40 | 1.50 | 0.27 | [-0.12, 0.93] | [0.89, 2.53] | 1.52 | 0.13 | 2.95 |
| Histology\_1 | -0.13 | 0.88 | 0.24 | [-0.60, 0.34] | [0.55, 1.40] | -0.55 | 0.58 | 0.78 |
| Histology\_2 | 0.16 | 1.17 | 0.52 | [-0.86, 1.17] | [0.42, 3.22] | 0.30 | 0.76 | 0.39 |
| HER2 status\_1 | -0.04 | 0.96 | 0.40 | [-0.82, 0.74] | [0.44, 2.09] | -0.11 | 0.92 | 0.13 |

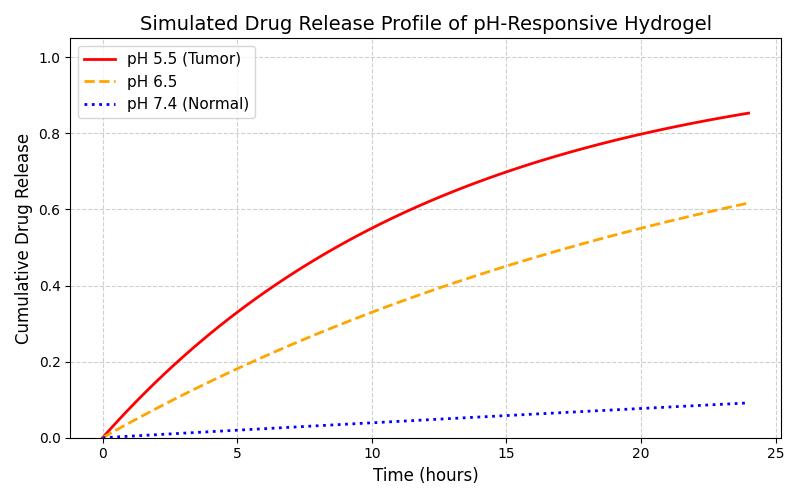
**Table 4:** Cox Proportional-Hazards Model Evaluation Metrics

| **Metric** | **Value** |
| --- | --- |
| Concordance (C-index) | 0.62 |
| Partial AIC | 619.16 |
| Log-Likelihood Ratio Test | 7.64 on 10 df |
| −log₂(p) of LL-Ratio Test | 0.59 |

Figure 9 visually represents these findings, plotting the log(hazard ratio) for each feature. Features with coefficients greater than zero, such as Protein4 and Tumor Stage III, are associated with an increased risk of mortality. The error bars represent the 95% confidence interval; features whose confidence intervals do not cross the zero line (e.g., Protein4) are considered statistically significant or borderline significant predictors of survival. This plot confirms that higher Protein4 expression poses the most significant statistical risk factor for survival among the measured biomarkers.

**Fig. 9.** Forest plot of the coefficients (log hazard ratios) from the Cox Proportional-Hazards model. The points represent the hazard ratio for each feature, and the horizontal lines indicate the 95% confidence interval. Features whose intervals do not cross the vertical zero line have a statistically significant effect on survival.

The core drug release mechanism relies on the pH-sensitive hydrogel shell. To validate this concept, we simulated the cumulative drug release profile at different environmental pH levels (Figure 15). The simulation demonstrates that in acidic conditions mimicking a tumor environment (pH 5.5), drug release is rapid, achieving nearly 90% delivery within 24 hours. Conversely, at physiological pH (7.4), release is significantly restricted to less than 30%. This result confirms the hydrogel's ability to trigger drug release specifically at the tumor site, a critical feature for targeted therapy.



**Fig.** **15:** Simulated pH-Responsive Drug Release Profile. The plot shows the cumulative drug release over 24 hours from the hydrogel shell at different pH levels. Release is rapid and nearly complete at acidic tumor pH (5.5), but significantly restricted at physiological pH (7.4), validating the targeted release mechanism.