NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Department of Computer Science and Engineering

Predicting Breast Cancer Survival: An Approach using Deep Learning and Machine Learning Techniques

PRESENTED BY

Chennupalli Chandrika Tirumala (22471A05L7)

Kodavati Jayamma (22471A05M7)

Parlapalli Haseena (23475A0501)

Under the Guidance of,

Soma Sekhar Kolisetty M.Tech,

Assistant Professor,

Department of Computer Science and Engineering,

Narasaraopeta Engineering College (Autonomous),

Narasaraopet- 522 601.

OUTLINE

1. Abstract
2. Introduction
3. Literature Survey
4. Research Gaps
5. Problem Statement
6. Objectives
7. Block Diagram / Flow Diagram
8. Methodology
9. Implementation
10. Results and Analysis
11. User Interfaces
12. Conclusion & Future Scope
13. References
14. Question and Answers
15. Acknowledgements

ABSTRACT

This project predicts **breast cancer survival** using the **METABRIC clinical and genomic dataset** by comparing **CoxPH, RFS, DeepSurv, DeepHit**, and the proposed **SA-DGNet** model. The **CoxPH** model achieves the highest **accuracy (C-index = 0.87)**, showing strong performance on **structured clinical data**, while **RFS** also performs well with a **C-index of 0.82** and provides good **interpretability**. The proposed **SA-DGNet** integrates **gated layers** and **self-attention mechanisms** for **dynamic and personalized risk prediction**, achieving a **C-index of 0.75**. Although **deep learning models** offer higher **flexibility**, traditional models such as **CoxPH** and **RFS** remain more effective on **limited datasets**, highlighting the future potential of **hybrid models like SA-DGNet** for **clinical decision support**.

INTRODUCTION

Brief Introduction to the Project

- This project predicts **breast cancer survival** using the **METABRIC dataset.**
- It aims to evaluate models like **CoxPH, RFS, DeepSurv, DeepHit, and SA-DGNet** for accurate survival prediction.
- The **SA-DGNet model** improves prediction and understanding using **self-attention and gating**.

Motivation Behind the Project

- Early prediction helps doctors give better treatment and care for breast cancer patients.
- Existing models are hard to understand and not suitable for complex medical data.
- **Using self-attention and gating methods** can make predictions more accurate and personalized for each patient.

Importance and Relevance

- Helps doctors find high-risk breast cancer patients early and give better treatment.
- Supports personalized and data-based healthcare decisions for patients.

LITERATURE SURVEY

No	Title	Author	Journal Name & Year	Methodology Adapted	Key Findings	Gaps
1	*Bio Fusion Net Deep Learning-Based Survival Risk Stratification in ER⁺ Breast Cancer Through Multi feature and Multimodal Data Fusion*	Raktim Kumar Mondol et al., 2024, IEEE Journal of Biomedical and Health Informatics.	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac.in/document/10568932	DINO, MoCoV3, Autoencoder (VAE), Self-Attention, Transformer Encoder, and Weighted Cox Loss	Proposed BioFusionNet model integrating imaging, genetic, and clinical data with weighted Cox loss, achieving superior survival prediction	Existing models struggle with effective multimodal data integration and imbalanced survival data.
2	*Deep-CSA: Deep Contrastive Learning for Dynamic Survival Analysis With Competing Risks*	Caogen Hong, Fan Yi, Zhengxing Huang, 2022, IEEE Journal of Biomedical and Health Informatics.	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac.in/document/9756287	LSTM encoder, Time-LSTM decoder, contrastive learning	Proposed Deep-CSA model using contrastive learning to improve survival prediction with competing risks on longitudinal data	Existing models fail to learn dynamic risk-specific representations from censored and sequential data
3	*Efficient Training of Probabilistic Neural Networks for Survival Analysis*	Christian Marius Lillelund, Martin Magris, Christian Fischer Pedersen, 2024, IEEE	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac.in/document/10568314	Variational Inference (VI), Monte Carlo Dropout (MCD), Spectral-normalized Neural Gaussian Process (SNGP)	MCD and SNGP match or outperform VI in prediction and calibration with lower computational cost	Prior works mostly rely on VI, with little exploration of alternative efficient Bayesian methods in survival analysis
4	*Tab-Coxn: An Interpretable Deep Survival Analysis Model for Patients With Nasopharyngeal Carcinoma Based on TabNet.*	Huamei Qi, Yuxuan Hu, Ruohao Fan, Lei Deng, 2024, IEEE Journal of Biomedical and Health Informatics	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac.in/document/10529284	TabNet integrated with Cox proportional hazards model (Tab-Cox) compared with DeepSurv, DeepHit, RSF	Tab-Cox improves prediction accuracy and interpretability on , tabular survival data and identifies clinically relevant risk factors	Existing deep survival models struggle with tabular data and lack interpretability for clinical use
5	*Using Bayesian Neural Networks to Select Features and Compute Credible*	Shi-ang Qi, Neeraj Kumar, Ruchika Verma, Jian-Yi Xu, 2023	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac.in/document/10158019	Bayesian Neural Networks (BNNs) with spike-and-slab and horseshoe priors	Proposed BNN-ISD models enable feature selection and credible interval computation	Traditional ISD models lack principled feature selection and uncertainty quantification mechanisms, limiting

LITERATURE SURVEY

No	Title	Author	Journal Name & Year	Methodology Adapted	Key Findings	Gaps
6	*Deep Survival Analysis With Latent Clustering and Contrastive Learning*	Chang Cui et al., 2024, IEEE Journal of Biomedical and Health Informatics	https://ieeexplore-ieee org.egateway.chennai.vit.ac .in/document/10423126	DeepHit backbone, latent clustering, contrastive loss	Jointly optimizes clustering and survival prediction, improving C-index/IBS	Existing models ignore inter-instance relationships and underutilize censored data
7	*Using Bayesian Neural Networks to Select Features and Compute Credible Intervals for Personalized Survival Prediction*	Shi-ang Qi, Neeraj Kumar, Ruchika Verma, Jian-Yi Xu, 2023, IEEE Transactions on Biomedical Engineering	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac .in/document/10158019	Bayesian Neural Networks (BNNs) with spike-and-slab and horseshoe priors, using CoxPH and MTLR survival models	Proposed BNN-ISD models enable feature selection and credible interval computation with state-of-the-art survival prediction accuracy	Traditional ISD models lack principled feature selection and uncertainty quantification mechanisms, limiting clinical trust and interpretability
8	*RESurv: A Deep Survival Analysis Model to Reveal Population Heterogeneity by Individual Risk*	Qiguang Zheng, Qifan Shen, Xin Su, Kuo Yang, Zixin Shu, 2022,IEEE International Conference on Bioinformatics and Biomedicine (BIBM)	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac . in/document/9995313	Multi-Layer Perceptron (MLP) and Gated Recurrent Unit (GRU) integrated into the RESurv framework	RESurv achieves state-of-theart survival prediction while revealing individual-level risk heterogeneity using counterfactual analysis	Existing deep survival models lack interpretability and do not effectively uncover patient-level heterogeneity
9	*HitBoost: Survival Analysis via a Multi-Output Gradient Boosting Decision Tree Method*	Pei Liu, Bo Fu, Simon X. Yang, 2019, IEEE Access	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac. in/document/8700177	HitBoost using MultiOutput Gradient Boosting Decision Trees (GBDT), implemented with XGBoost	HitBoost outperforms classical survival models in prediction and feature importance analysis without assuming hazard function forms	Traditional models rely on assumptions and deep models lack interpretability and require intensive training
10	Effective Strategies and Techniques Used for Pulmonary Carcinoma Survival Analysis	Poornima G & Anand L, 2024, IEEE Conference Proceedings	https://ieeexplore-ieeeorg.egateway.chennai.vit.ac. in/document/10576100	Cox model, KaplanMeier, clustering, deep learning classifiers	Surveys survival analysis and highlights cancer-specific model variations	Limited application of clustering and deep models in lung cancer survival

RESEARCH GAPS

1. Limited Feature Integration

Most studies use either clinical or genomic data alone. The combined use of both types of data is still limited.

2. Less Use of Deep Learning in Medical Data

Advanced models like BERT are used for text data but rarely applied to medical survival prediction.

3. Poor Generalization of Models

Models trained on one dataset do not perform well on other datasets because of differences in patient data.

4. Data Imbalance Problem

Many datasets have more records for survivors than non-survivors, which reduces prediction accuracy.

5. Lack of Explainable Models

Deep learning models give good results but are hard to understand. More focus is needed on explainable AI for medical use.

PROBLEM STATEMENT

1.Reasearch Gap Problem

Existing studies mostly use **only clinical or only genomic data**, which limits the ability to capture complete patient risk patterns.

Solution:

In the proposed work, **both clinical and genomic features from the METABRIC dataset** are jointly used and evaluated through **CoxPH, RFS and the proposed SA-DGNet**, enabling richer and more reliable survival prediction.

2.Reasearch Gap Problem

Most medical survival studies still rely on traditional models, and advanced deep models are **not effectively explored for dynamic survival modeling**.

Solution:

This work introduces a deep model called **SA-DGNet**, which uses **gated layers and self-attention mechanisms** to learn complex and time-dependent survival patterns.

PROBLEM STATEMENT

3. Research Gap Problem

Survival models often fail to perform consistently because of **heterogeneous and structured medical data**.

Solution:

The study benchmarks strong and stable classical models (**CoxPH and RFS**) together with **SA-DGNet**, providing a **robust and generalizable survival prediction framework** on the METABRIC dataset.

4. Research Gap Problem – Data Imbalance and Limited Dataset Size

Breast cancer datasets are often **imbalanced and small**, causing deep models to overfit and reducing prediction accuracy.

Solution:

The proposed framework emphasizes **reliable traditional models (CoxPH and RFS)** for stable learning on limited data, while using **SA-DGNet** to complement them for advanced pattern learning.

PROBLEM STATEMENT

5. Research Gap Problem

Deep learning models provide predictions but offer **low interpretability**, making them difficult for clinical adoption.

Solution:

The proposed approach combines **interpretable models (CoxPH and RFS)** with an attention-based deep model (**SA-DGNet**) that provides **explainable and time-aware risk estimation**, supporting practical clinical decision making.

OBJECTIVES

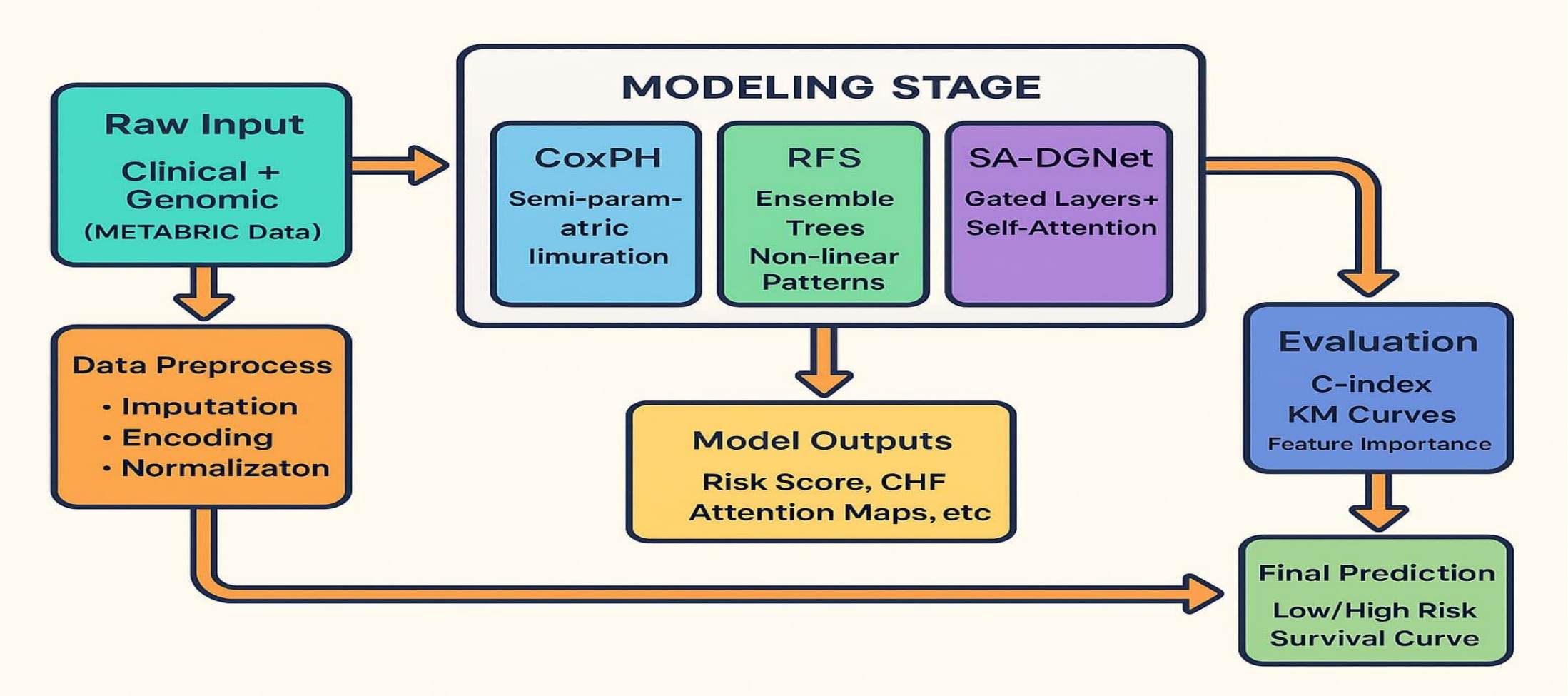
Main Objective:

- To develop an advanced Deep Learning and Machine Learning–based model for accurate and reliable **breast cancer survival prediction** using clinical and genomic data.

Specific Objectives:

- **To analyze** the limitations of existing breast cancer prediction techniques.
- **To preprocess** and clean the breast cancer dataset by handling missing values and class imbalance.
- **To extract** important clinical, demographic, and genetic features influencing patient survival.
- **To implement** Machine Learning algorithms for baseline survival prediction.
- **To design and develop** a Deep Learning model capable of learning complex nonlinear relationships among features.
- **To evaluate** the model using standard performance metrics such as Accuracy, Precision, Recall, F1-score, and AUC.
- **To compare** the proposed hybrid model with existing methods to demonstrate its **efficiency** and **robustness**.

BLOCK DIAGRAM OR FLOW DIAGRAM



METHODOLOGY

1.Dataset Selection:
Used METABRIC dataset containing clinical and genomic data from over 1900 breast cancer patients.
Preprocessing
Performed data cleaning and Min–Max normalization.
Handled missing values (median/mode) and applied one-hot encoding for categorical variables.

2. Feature Analysis & Grouping:
Identified key risk and protective features via CoxPH hazard ratios and RFS permutation importance.
• Grouped patients into High-Risk and Low-Risk categories using model-predicted risk scores and Kaplan–Meier survival curves.

METHODOLOGY

3. Model Framework:

- **Cox Proportional Hazards (CoxPH):** Linear statistical model estimating hazard ratios.
- **Random Forest Survival (RFS):** Tree-based ensemble model for nonlinear interactions.
- **Proposed SA-DGNet:** Combines **Gated Neural Network** and **Self-Attention** for temporal feature learning.

4. Evaluation Metrics:

Concordance Index (C-index) used to assess ranking accuracy.
Mean Absolute Error (MAE) measured prediction deviation.
Monitored training curves for loss and C-index evolution to ensure stability and generalization.
Best results achieved by **CoxPH (C-index = 0.87)** and **RFS (C-index = 0.82)**.

IMPLEMENTATION

Implementation:
The project was done using **Python** in **Google Colab**. Performed **data cleaning**, **normalization**, and **encoding**. Used the **METABRIC dataset** with data of 1,900+ breast cancer patients. Training models: CPH, RFS, DeepSurv, SA-DGNet.

Software Specs:
OS: Colab (Linux)
Frameworks: PyTorch, Scikit-learn, Pycox, Lifelines
Tools: Pandas, NumPy, Matplotlib, Seaborn

Hardware Specs:
GPU: NVIDIA Tesla T4
RAM: 8GB | Storage: Around 10 GB
Processor: Intel Core i5/i7

Challenges & Solutions:
Missing data: Fixed using median/mode values.
Overfitting → Used dropout and early stopping.
Model explainability → Added self-attention layer.
Slow training → GPU acceleration

RESULTS & ANALYSIS

2. Training and Validation Loss Analysis

The graph shows how the **training** and **validation losses** change over **25 epochs** during model training.

•The **training loss** (blue line) **decreases steadily**, indicating that the model is **learning well** from the training data.

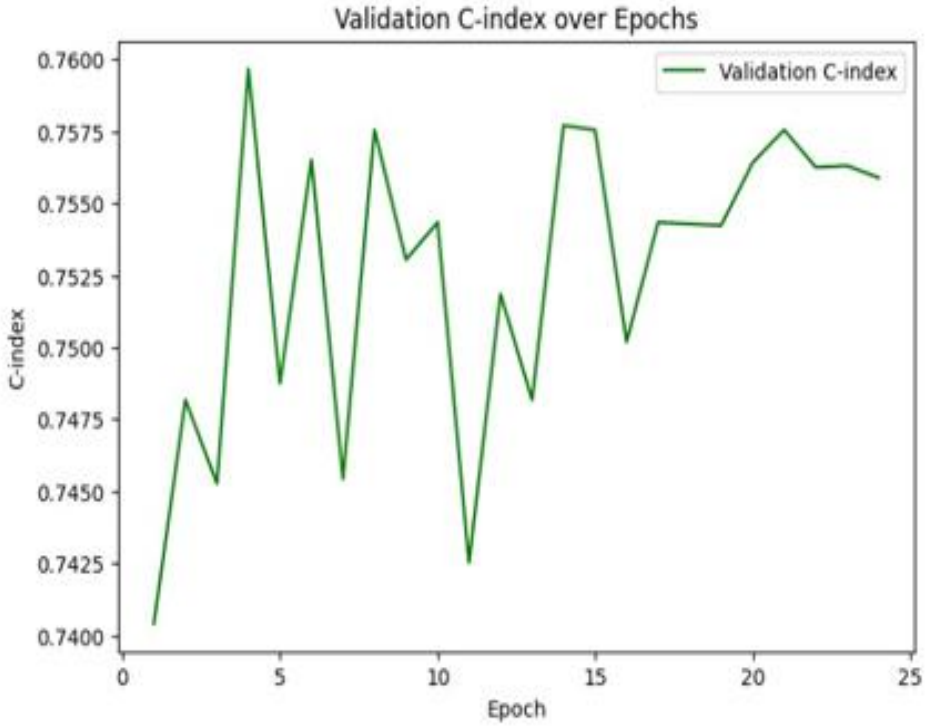
•The **gap between the two curves** suggests **slight overfitting** — the model performs better on training data than on validation data.

•Overall, the model shows **good convergence** but may require **regularization, dropout, or early stopping** to reduce overfitting and improve validation performance.

• 📈 ***Visual Insight***:

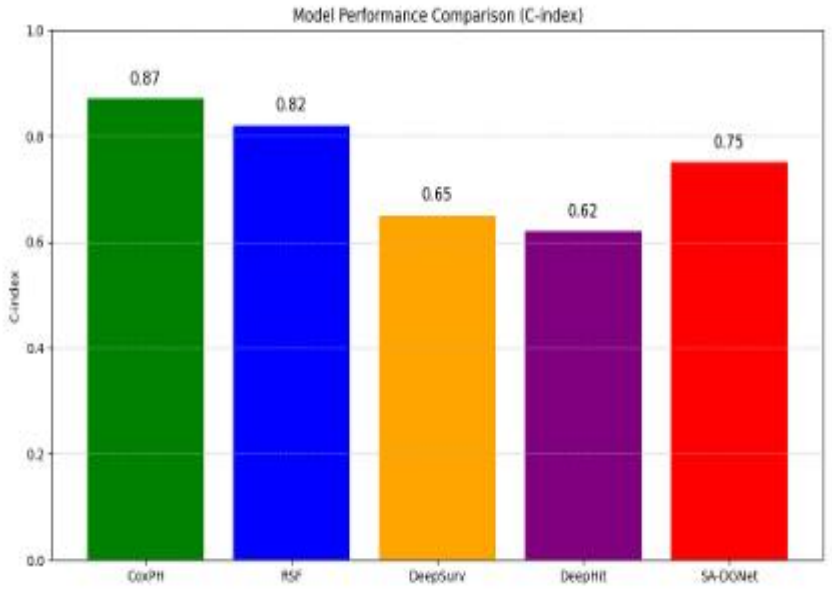
•The C-index curve shows how the model's **prediction accuracy improves** with each training epoch.

RESULTS & ANALYSIS



- The upward trend indicates that the model's **discriminative ability improved** with training.
- **X-axis:** Epoch
- **Y-axis:** C-index(Model Performance Score)
- The **C-index** measures how well the model predicts survival rankings — a **higher value means better performance**.
- At the start, the C-index fluctuates as the model learns, but later it becomes **more stable around 0.75–0.76**, showing **consistent prediction accuracy**.
- Overall, the results show **good generalization** on validation data, proving that the model effectively learned the survival patterns.

RESULTS & ANALYSIS



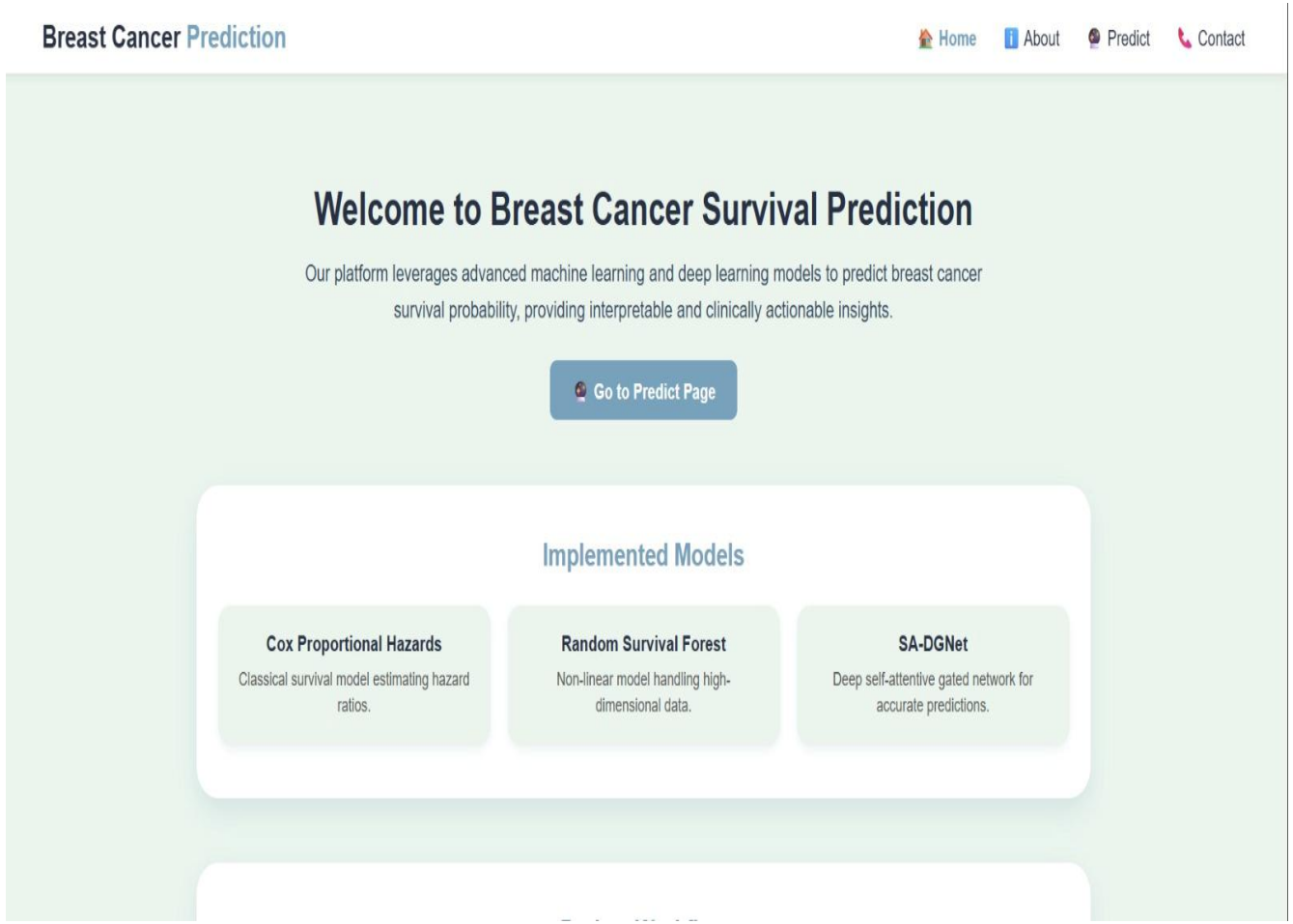
Model Performance(C-index)

- The graph compares how well each model predicts survival.
- The bar chart compares the **C-index values** of different survival prediction models.
- The **Model Performance Comparison** graph highlights that **CPH** achieved the **best performance (0.87)**, followed by **RFS (0.82)**, while the proposed **SA-DGNet (0.75)** showed good potential for improvement.

✅ *Conclusion:*

•The visual results show that the models learned well with stable performance. CPH gave the best accuracy, while SA-DGNet showed good potential. Overall, the models performed effectively for survival prediction.

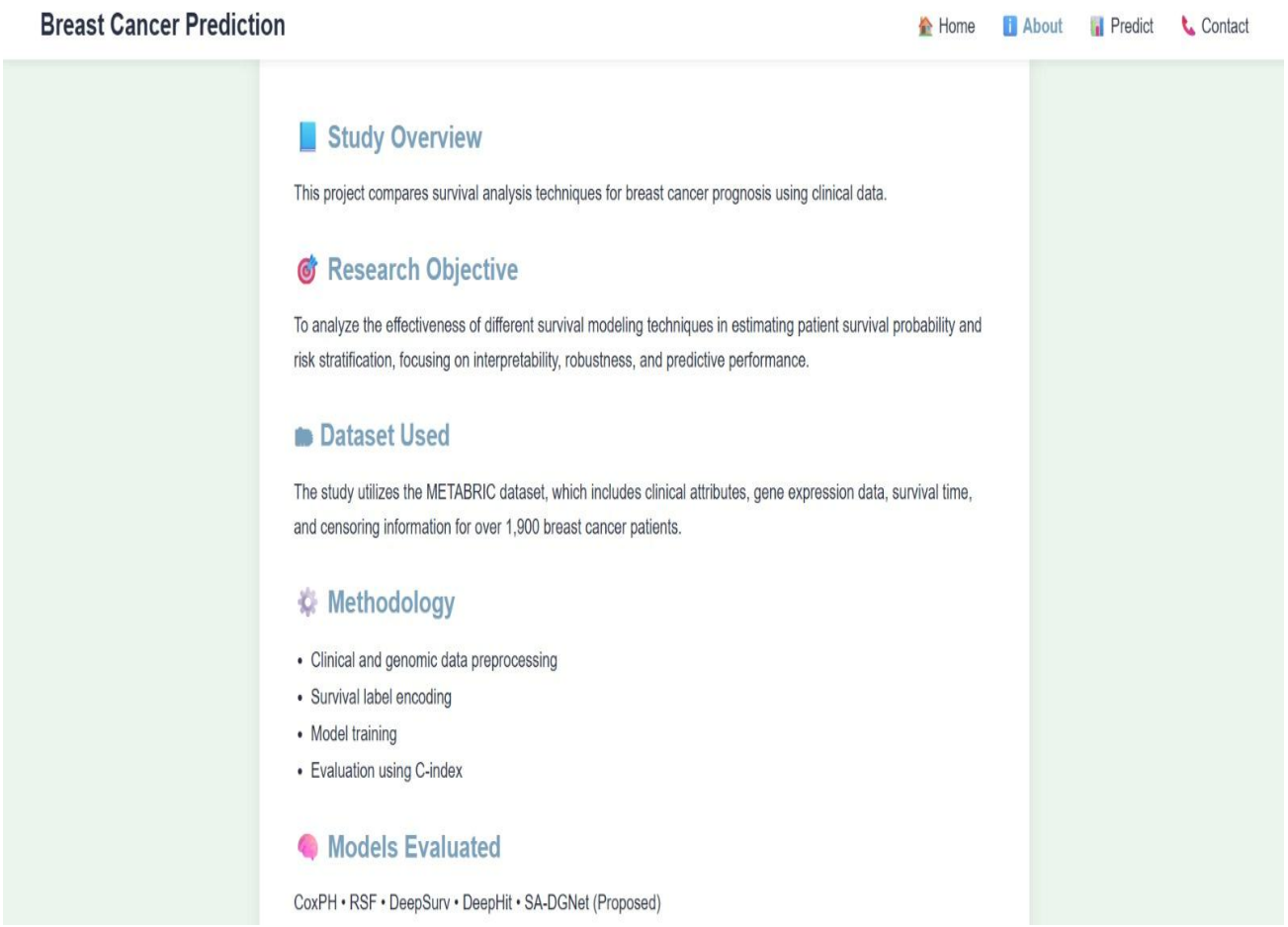
OUTPUT SCREENS



Home Page

- The Home page introduces the **Breast Cancer Survival Prediction system**.
- It explains that the system uses **machine learning and deep learning** to predict survival.
- A **Go to Predict Page** button is provided for prediction access..
- An **Implemented Models** section is displayed.
- The models shown are **Cox Proportional Hazards, Random Survival Forest and SA-DGNet**.
- Each model is briefly described to show its purpose in prediction.

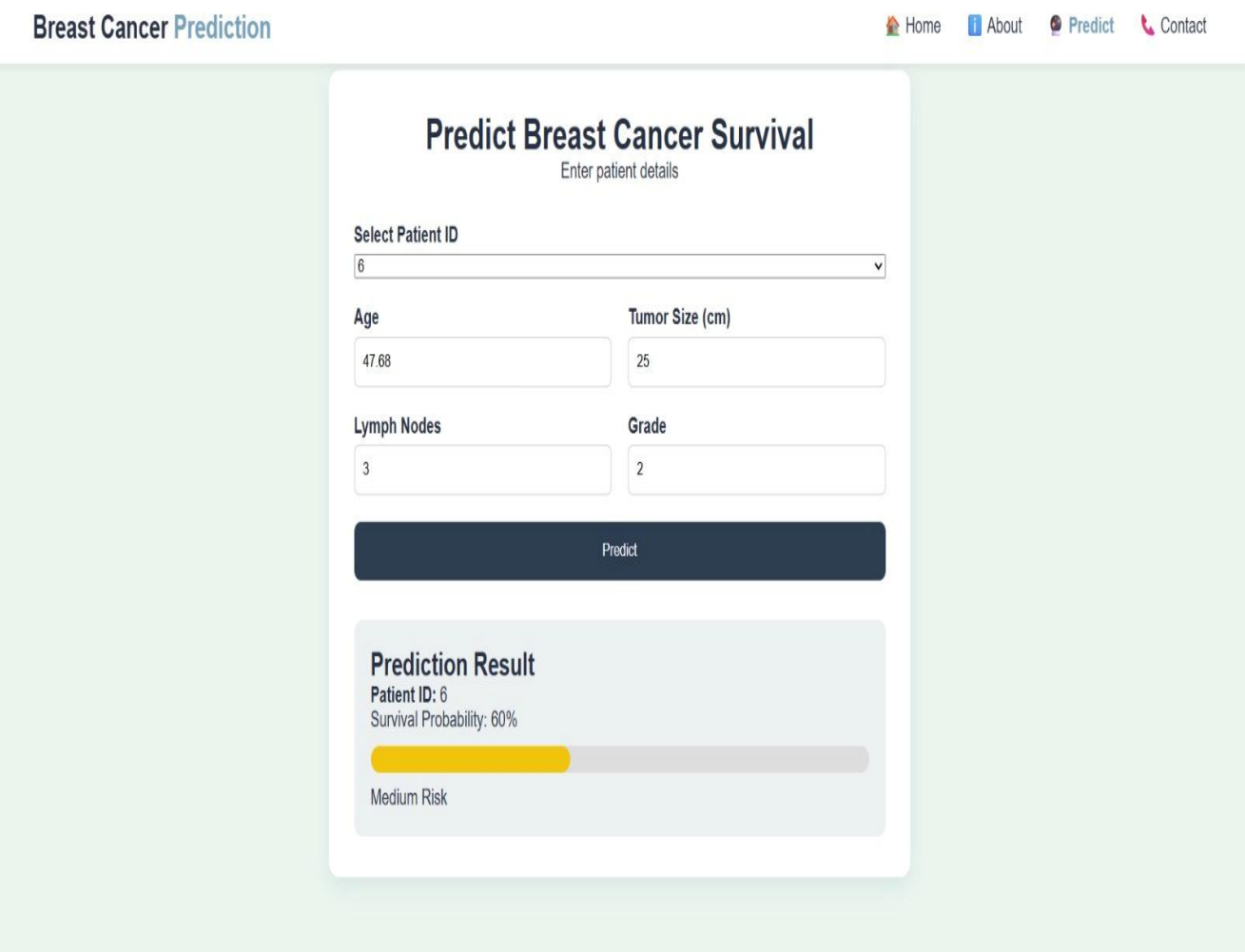
OUTPUT SCREENS



About Page

- The page gives a **study overview** of breast cancer survival prediction using survival analysis techniques.
- It clearly states the **research objective** of analyzing effectiveness, interpretability, and robustness of survival models.
- The **METABRIC dataset** is used, containing clinical and gene expression data of more than 1,900 patients.
- Multiple **survival models are evaluated**, including both classical and deep learning approaches.
- The evaluated models are **CoxPH, RSF, DeepSurv, DeepHit, and the proposed SA-DGNet**.

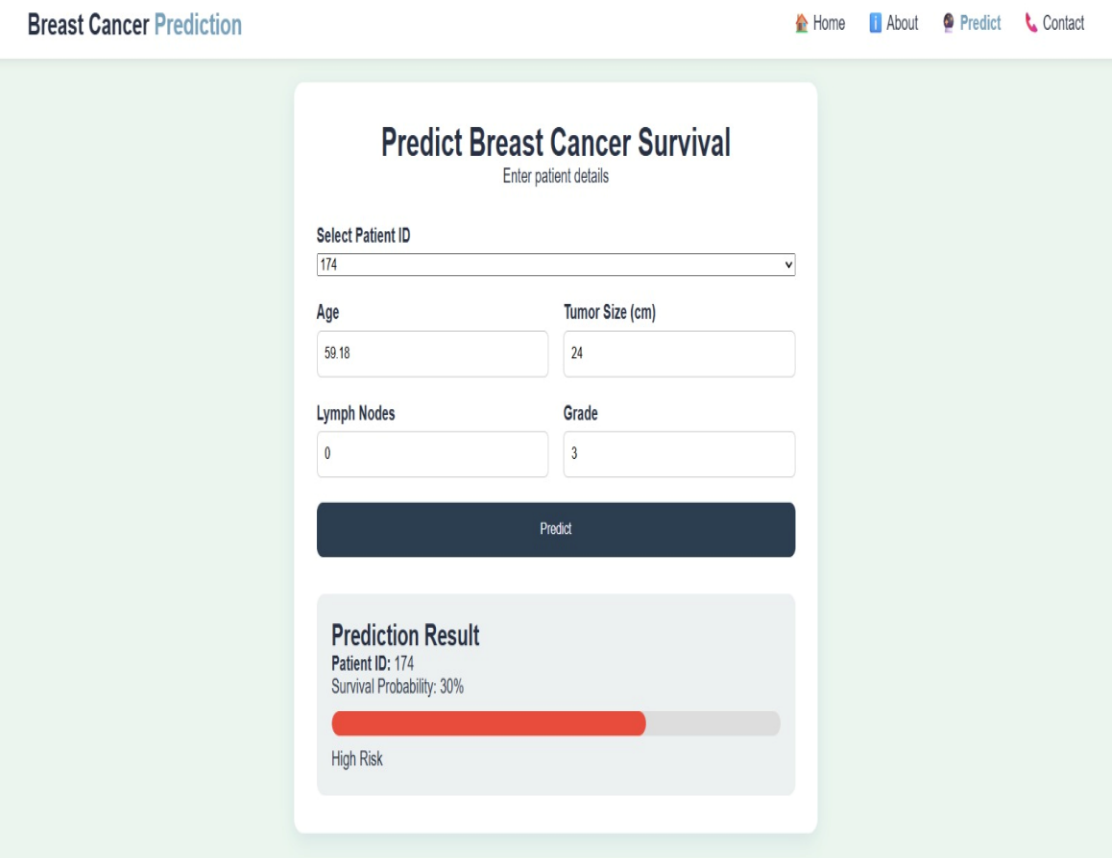
OUTPUT SCREENS



Predict Page

- The Predict page allows users to **enter and modify patient clinical details** to estimate breast cancer survival.
- Users can select a **Patient ID**, which automatically loads the corresponding patient information.
- The input fields include important clinical factors such as **Age, Tumor Size, Lymph Nodes, and Tumor Grade**.
- After entering the details, the user clicks the **Predict** button to generate the prediction.
- The system displays the **survival probability percentage** for the selected patient.
- It also shows the patient's **risk category** along with a **visual progress bar**

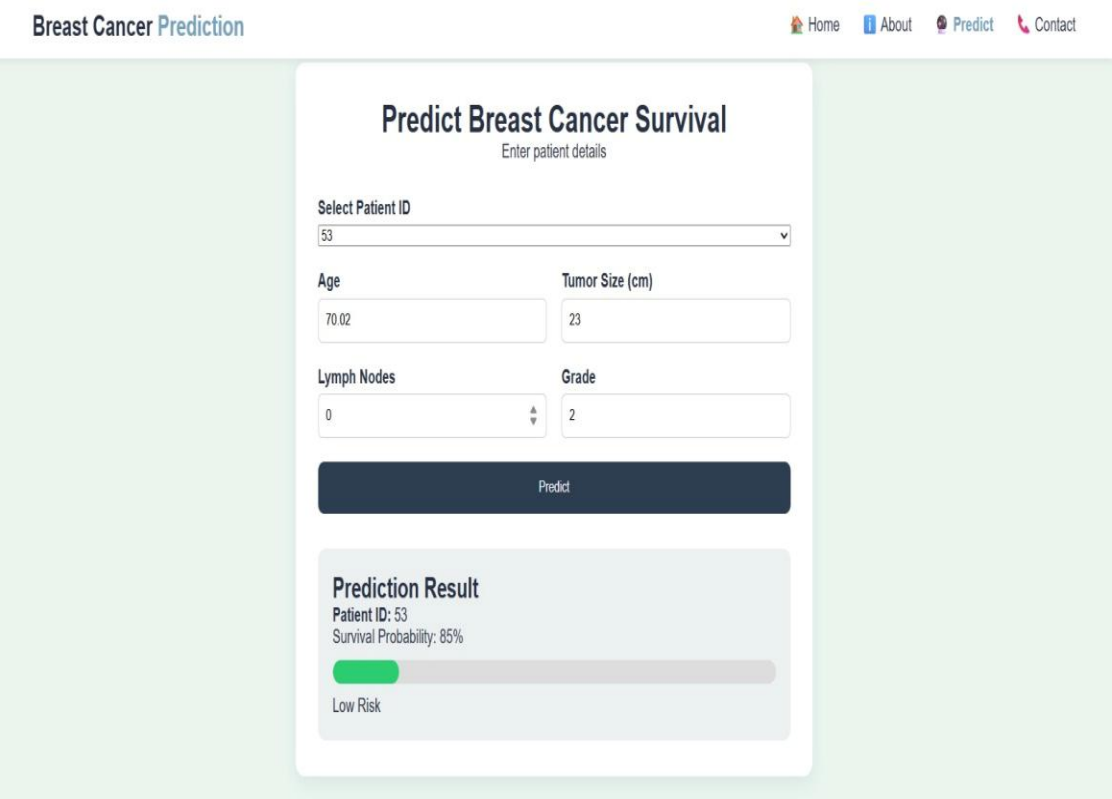
OUTPUT SCREENS



High-Risk Result:

- The system classifies the patient as **High Risk** when the predicted survival probability is low.
- A **High Risk label** is clearly displayed for the selected Patient ID.
- The screen shows a **low survival percentage** for the patient.
- A **red / warning progress bar** visually represents the high-risk level.
- High-risk status is determined based on clinical inputs such as **age, tumor size, lymph nodes and tumor grade**.
- This result helps doctors **prioritize intensive treatment and closer monitoring** for the patient.

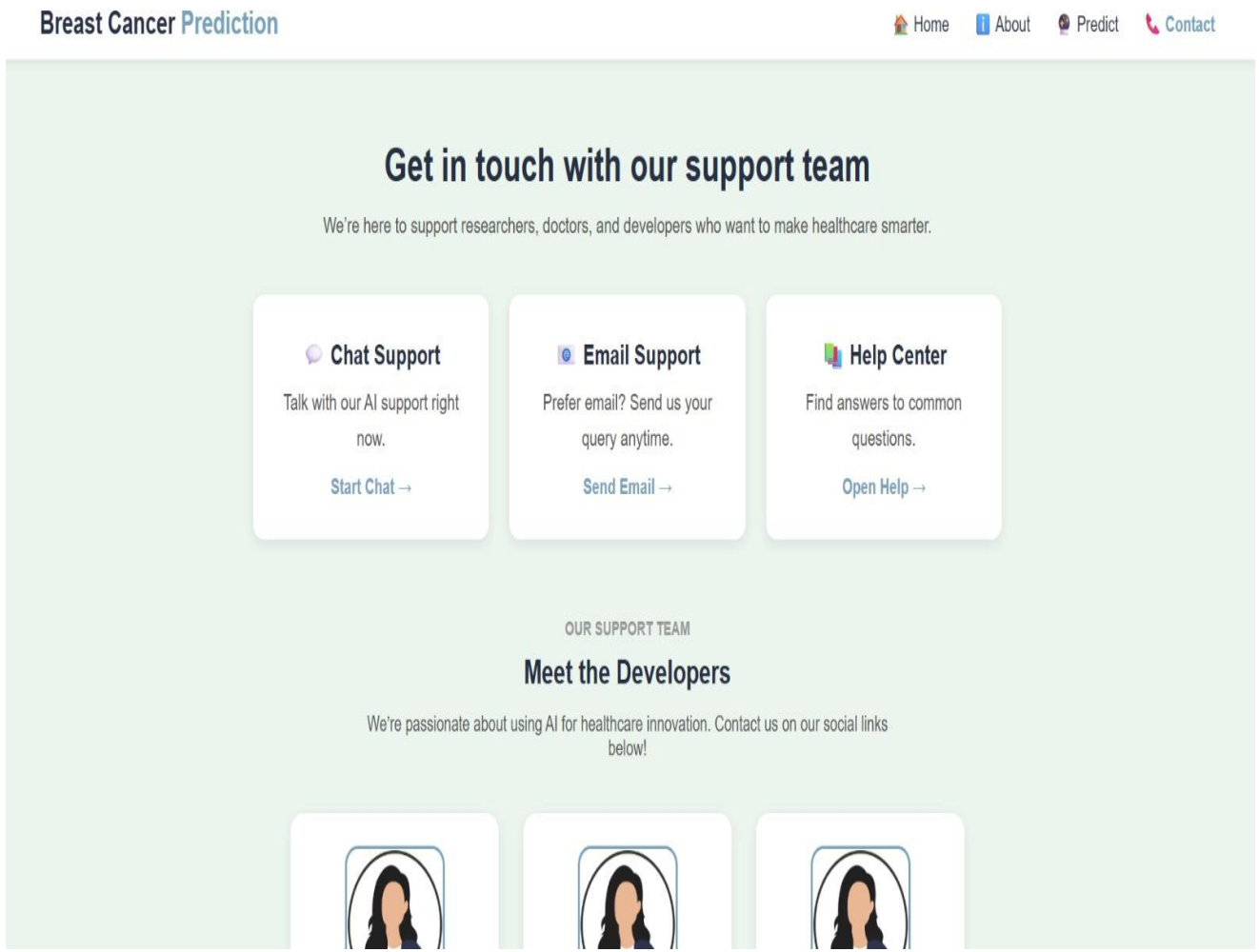
OUTPUT SCREENS



Low-Risk Result:

- The system classifies the patient as **Low Risk** when the predicted survival probability is high.
- A **Low Risk label** is clearly displayed for the selected Patient ID.
- The screen shows a **high survival probability percentage** for the patient.
- A **green / safe progress bar** visually represents the low-risk level.
- Low-risk status is determined using clinical inputs such as **age, tumor size, lymph nodes and tumor grade**.
- This result helps doctors **continue standard treatment and routine follow-up** for the patient.

OUTPUT SCREENS



Contact Page

- The Contact page allows users to **get in touch with the support team** for assistance related to the system.
- It provides **Chat Support** for instant help through an AI-based chat option.
- An **Email Support** option is available to send queries at any time.
- A **Help Center** section is provided to find answers to common questions.
- The page also introduces the **development team** under the "Meet the Developers" section.
- It encourages users to connect with the developers through **social links** for collaboration and support.

CONCLUSION and FUTURE SCOPE

Key Findings

- CPH gave the best accuracy (c-index=0.87).
- **RFS** also showed strong and stable results.
- SA-DGNet model added self-attention for better understanding of data patterns.
- Deep models helped improve interpretability of predictions.

Significance of Results

- Shows that **traditional models** still perform well on medical data.
- **Deep learning** can find hidden and complex patterns.
- Useful for **early prediction** and **better treatment planning** for patients.
- The study contributes to developing **AI-driven clinical decision-support systems** for personalized treatment planning.

CONCLUSION and FUTURE SCOPE

Future Developments

- Integrate multimodal data(genomic,imaging, and clinical).
- Implement transfer learning for improved generalization on smaller medical datasets.
- Develop a user-friendly web or mobile application for real-time survival prediction.
- Enhance interpretability using attention heatmaps and explainable AI methods.

Limitations

- Limited dataset size (METABRIC) may restrict deep model learning.
- Deep models require **high computational resources** and longer training times.
- Results depend on **data quality** and **missing value handling**.
- May not directly generalize to other cancer types without retraining.

REFERENCES

[1] Hong et al. (2022):
Proposed *Deep-CSA*, a contrastive learning model for dynamic survival analysis with competing risks, improving patient risk differentiation.
[2] Lillelund et al. (2024):
Developed efficient probabilistic neural networks for survival analysis, reducing training complexity in biomedical prediction.
[3] Qi et al. (2024):
Introduced *Tab-Cox*, an interpretable deep survival analysis model for cancer prognosis using tabular patient data.
[4] Chi et al. (2021):
Presented a deep semi-supervised multitask learning framework for survival analysis, integrating limited-labeled medical data.
[5] Cui et al. (2024):
Proposed latent clustering and contrastive learning for deep survival analysis, enhancing subgroup identification.
[6] Qi et al. (2023):
Applied Bayesian neural networks for personalized survival prediction, improving uncertainty estimation in medical outcomes.
[7] Zheng et al. (2022):
Designed *RESurv*, a deep survival model for population heterogeneity through individualized risk learning.

REFERENCES

[8] Chen et al. (2023):
Proposed transformer-based survival models for cancer prognosis, leveraging self-attention to capture complex dependencies.
[9] Zhou and Zhang (2023):
Introduced a hybrid neural network for clinical time-to-event prediction, combining statistical and deep features.
[10] Poornima and A. L. (2024):
Reviewed effective techniques for pulmonary carcinoma survival analysis, comparing classical and deep methods.
[11] Fotso (2018):
Developed a multi-task deep neural network for survival analysis, pioneering deep architectures for censored data.
[12] Kvamme and Borgan (2019):
Demonstrated continuous and discrete-time neural network approaches for survival prediction.
[13] Lee et al. (2020):
Proposed *DeepHit*, a deep learning model for competing-risk survival analysis with discrete-time estimation.
[14] Katzman et al. (2018):
Introduced *DeepSurv*, a Cox-based deep neural network for personalized treatment recommendation.
[15] Lucken et al. (2021):
Designed recurrent deep models for survival analysis with longitudinal data, improving time-dependent predictions.

QUESTIONS and ANSWERS

✨ Thank you for listening! ✨

We've shared our journey of building a **secure and intelligent Healthcare IoT system** using **Blockchain**, **Federated Learning**, and **Differential Privacy**.

Now it's **your turn** — we'd love to hear your **questions, thoughts, or suggestions**. 💬

- DG4

ACKNOWLEGEMENTS

Thank You!

We sincerely **thank the organizers, faculty, and audience** for giving us the opportunity to present our project —

"Predicting Breast Cancer Survival-An Approach Using Deep Learning and Machine Learning Techniques ."

Your attention, feedback, and questions are truly appreciated.

For any further discussions or collaboration inquiries, feel free to contact us:

Contact Information

Chennupalli Chandrika Tirumala – tirumalachennupalli@gmail.com

Soma Sekhar Kolisetty – sekhar.soma007@gmail.com

Kodavati Jayamma – k.jayamma879@gmail.com

Parlapalli Haseena– haseenaparlapalli17@gmail.com