Analysis of Utility-Kernal Based SVM for Survival Estimation of Breast Cancer

[1]Chittinni Pratyusha, [2]G.Praveen Babu
[1]*MCA Student, Department of Information Technology, Jawaharlal Nehru Technological University, India*
[2]*Associate Professor of CSE, Department of Information Technology, Jawaharlal Nehru Technological University, India*

Abstract: The advancement of medical research in cancer prognosis and diagnosis, particularly in breast cancer, has placed significant stress on oncologists due to the complexity and heterogeneity of the disease. To address this challenge, research focusing on breast cancer survival estimation has been proposed. This research aims to merge histological and genomic data to enhance prognosis accuracy, minimizing unnecessary treatment interventions and ensuring tailored patient care. By utilizing a spectrum of machine learning approaches, including SVM, Random Forest, and neural networks, a robust tool is developed for personalized breast cancer survival predictions. The tool empowers clinicians to make informed treatment decisions and optimize healthcare resource allocation efficiently. Moreover, the application of ensemble methods further enhances prediction accuracy, particularly through techniques like Voting Classifier. Additionally, extending the project to include a user-friendly frontend using the Flask framework allows for user testing and authentication, facilitating seamless interaction and practical application in clinical settings.

Index Terms—Breast cancer survival estimation, gene expression, copy number variation, histopathological whole slide images, utility kernel, support vector machine, machine learning, deep neural networks

1. INTRODUCTION

Breast cancer, characterized by the uncontrolled growth of breast cells, remains a significant global health concern, particularly for females. The intricacies of breast tissue composition and the complexities of cancer progression underscore the critical need for accurate prognosis and diagnosis tools. With advancements in medical research and technology, oncologists are increasingly relying on a combination of qualitative histological data and quantitative genomic information to predict clinical outcomes and tailor treatment strategies [1].

Traditional survival prediction models often face challenges in accurately predicting clinical outcomes due to the heterogeneity of breast cancer and the varied clinical responses observed among patients [2]. Recent advancements in medical imaging and next-generation sequencing technologies, such as METABRIC and The Cancer Genome Atlas (TCGA), have revolutionized the field by providing extensive genome-scale transcriptomic data for breast cancer research [3].

However, given the complexity of breast cancer and the variability in survival rates, newer approaches are needed to accurately classify patients as short-term or long-term survivors based on varying survival cutoffs [4].

Machine learning (ML) techniques have emerged as powerful tools in the development of survival prediction models for breast cancer. ML algorithms can automatically learn from data, handle complex interactions between variables, and excel at processing large volumes of medical data [5].

In this study, I aim to address the pressing need for accurate survival prediction models in breast cancer research. By integrating diverse datasets and employing advanced ML techniques, we seek to develop robust and personalized survival prediction tools. These tools have the potential to assist oncologists in making informed treatment decisions, optimize healthcare resource allocation, and ultimately improve patient outcomes.

2.LITERATURE SURVEY

This literature survey explores various studies and research papers that have contributed to our understanding of breast cancer prognosis, adherence challenges in patient care, genomic and transcriptomic architecture, and the role of big data

and machine learning in predicting survival outcomes.

Clark (1994) [1] discusses the necessity of prognostic factors for breast cancer, emphasizing the importance of identifying reliable indicators to guide treatment decisions and improve patient outcomes. This study highlights the complexities of breast cancer prognosis and the challenges associated with predicting disease progression.

Martin et al. (2005) [2] address the challenge of patient adherence in breast cancer treatment, emphasizing the importance of patient engagement and adherence to treatment plans for successful outcomes. This study underscores the need for effective communication between healthcare providers and patients to promote treatment adherence and improve clinical outcomes.

Curtis et al. (2012) [3] explore the genomic and transcriptomic architecture of breast tumors, revealing novel subgroups based on molecular characteristics. This study provides valuable insights into the heterogeneity of breast cancer and highlights the potential for personalized treatment approaches based on molecular profiling.

Tomczak et al. (2015) [4] provide a comprehensive review of The Cancer Genome Atlas (TCGA), emphasizing its role as an invaluable source of knowledge for cancer research. This study underscores the significance of large-scale genomic data in understanding the molecular basis of breast cancer and identifying potential therapeutic targets.

Delen et al. (2005) [5] compare three data mining methods for predicting breast cancer survivability, highlighting the efficacy of machine learning techniques in prognostic odelling. This study demonstrates the potential of data-driven approaches in predicting clinical outcomes and guiding treatment decisions in breast cancer.

Overall, the literature survey highlights the multidimensional nature of breast cancer prognosis and the importance of integrating diverse data sources and advanced analytics techniques to improve predictive accuracy and patient outcomes.

3.METHODLOGY

a) Proposed work:

The proposed breast cancer survival prediction system integrates advanced machine learning techniques with multimodal data sources to enhance prognostic accuracy. It utilizes a utility kernel for Support Vector Machines (SVM), incorporating gene expression data and histopathological images for improved performance compared to existing methods. Experimentation with Naive Bayes, Decision Trees, and Random Forests supplements the analysis. Deep learning methods, such as variational autoencoders and transfer learning with Convolutional Neural Networks (CNNs), further extend the model's capabilities.

As an extension, a voting classifier combining Random Forest, Support Vector Classifier, and Decision Tree models is implemented to leverage multi-modal data combinations for survival prediction at various intervals. Moreover, a Flask framework integrated with SQLite is developed to enable user signup and signin functionalities, facilitating user testing by allowing input submission and evaluation of the proposed models' performance.

b) System Architecture:

The system architecture begins with the input of the breast cancer dataset, comprising gene expression data and histopathological images. Data preprocessing involves cleaning, normalization, and feature extraction, followed by visualization to gain insights into the dataset's characteristics. The dataset is then split into training and testing sets for model development and evaluation.

Various machine learning algorithms, including Random Forest, Decision Tree, Naive Bayes, VGGNet, ResNet, and DenseNet, are employed for survival prediction.

The system architecture also includes a Flask framework integrated with SQLite for user signup, signin.

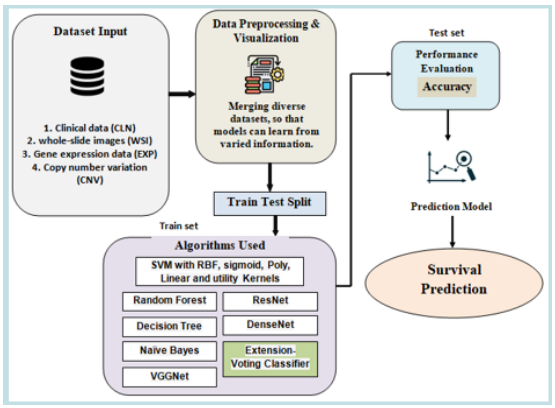

Fig 1 Proposed Architecture

c) Data processing

1. Data Sampler:

- Subsampling the long data to create a balanced dataset for training.

2. Define Utility Kernel:

- Definition of a utility kernel for Support Vector Machines (SVM), which enhances the SVM's ability to handle complex and non-linear relationships in the data.

3. Applying PCA:

- Principal Component Analysis (PCA) is applied to the dataset to reduce dimensionality while preserving 95% of the explained variance ratio, aiding in feature selection and model efficiency.

4. Exploring the Dataset:

- Analysis of different data components:

  - CLN: Clinical data

  - WSI: Whole slide images

  - EXP: Gene expression data

  - CNV: Copy number variation data

5. Combining the Data:

- Various combinations of data components are created for analysis and model training:

-CLN_EXP,CLN_CNV,CLN_WSI, EXP_CNV, EXP_WSI,CNV_WSI,CLN_EXP_CNV,CLN_EXP_WSI,EXP_CNV_WSI,CNV_CLN_WSI, multimodal (WSI, CLN, CNV, EXP).

6. Visualization:

- Data visualization using Seaborn and Matplotlib libraries to gain insights into the dataset's distributions, correlations, and patterns.

d) TRAINING AND TESTING

The dataset is divided into two subsets: a training set and a testing set. The training set comprises a balanced representation of breast cancer samples, including both short-term and long-term survivors, ensuring the model learns from a diverse range of instances. The SVM model is trained on this dataset using the utility kernel, which enhances the SVM's ability to capture complex relationships within the data.

Once the model is trained, it is evaluated on the testing set to assess its performance and generalization ability. The testing set contains unseen samples that were not used during training, allowing for an unbiased evaluation of the model's predictive accuracy. Performance metrics such as accuracy, precision, recall, and F1-score are calculated to gauge the model's effectiveness in accurately predicting breast cancer survival outcomes.

e) ALGORITHMS:

CLN - SVM – rbf

The RBF kernel allows the SVM to capture non-linear relationships in the CLN data, enhancing its ability to discern patterns associated with different survival outcomes. By leveraging CLN-SVM-RBF, the project aims to develop a robust predictive model that can assist oncologists in making informed treatment decisions and improving patient care in breast cancer management.

CLN - SVM – Utility

The utility kernel enhances the SVM's ability to handle complex and non-linear relationships within the CLN data, thereby improving its predictive accuracy. By leveraging CLN-SVM-Utility, the project aims to develop a robust predictive model that assists oncologists in making informed treatment decisions and enhances patient care in the field of breast cancer management.

CLN - SVM – Linear

The linear kernel facilitates the SVM in delineating linear relationships within the CLN data, aiding in the classification of survival outcomes. By leveraging CLN-SVM-Linear, the project aims to develop an effective predictive model that assists oncologists in making informed treatment decisions and enhances patient care in breast cancer management by providing accurate survival estimations based on clinical data.

CLN - SVM - Sigmoid

The sigmoid kernel allows the SVM to capture non-linear relationships within the CLN data, enabling it

to discern intricate patterns associated with different survival outcomes. By leveraging CLN-SVM-Sigmoid, the project aims to develop a robust predictive model that assists oncologists in making informed treatment decisions and enhances patient care by providing accurate survival estimations based on clinical data.

CLN-WSI-EXP-CNV – NB

CLN-WSI-EXP-CNV-NB represents a Naive Bayes (NB) classifier trained on a combination of clinical (CLN), whole slide images (WSI), gene expression (EXP), and copy number variation (CNV) data. By integrating information from diverse data types, including clinical attributes, imaging data, and molecular profiles, the CLN-WSI-EXP-CNV-NB model captures comprehensive insights into the disease. Through this approach, the project aims to develop a robust predictive model that assists oncologists in making informed treatment decisions and enhances patient care in breast cancer management.

CLN-WSI-EXP-CNV- voting

CLN-WSI-EXP-CNV trained on a combination of clinical (CLN), whole slide images (WSI), gene expression (EXP), and copy number variation (CNV) data. In the project, this model combines the predictions of multiple individual classifiers, such as Random Forest, Support Vector Machine, and Decision Tree, to make final predictions. By aggregating the results of diverse models trained on different data modalities, CLN-WSI-EXP-CNV-Voting aims to improve predictive accuracy and robustness.

CLN-WSI-EXP-CNV – VGGNets

CLN-WSI-EXP-CNV-VGGNets refers to a deep learning model utilizing VGGNet architecture trained on a combination of clinical (CLN), whole slide images (WSI), gene expression (EXP), and copy number variation (CNV) data. VGGNet, known for its deep architecture and strong feature extraction capabilities, enhances the model's ability to capture intricate patterns and relationships within the multi-modal data. By leveraging CLN-WSI-EXP-CNV-VGGNets, the project aims to develop a highly accurate predictive model.

CLN-WSI-EXP-CNV – ResNets

CLN-WSI-EXP-CNV-ResNets refers to a deep learning model utilizing ResNet architecture trained on a combination of clinical (CLN), whole slide images (WSI), gene expression (EXP), and copy number variation (CNV) data. ResNet, renowned for its deep architecture and residual connections, enhances the model's ability to capture intricate patterns and relationships within the multi-modal data. By leveraging CLN-WSI-EXP-CNV-ResNets, the project aims to develop a highly accurate predictive model.

4.EXPERIMENTAL RESULTS

Accuracy: The accuracy of a test is its ability to differentiate the patient and healthy cases correctly. To estimate the accuracy of a test, we should calculate the proportion of true positive and true negative in all evaluated cases. Mathematically, this can be stated as:

Accuracy = TP + TN TP + TN + FP + FN.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

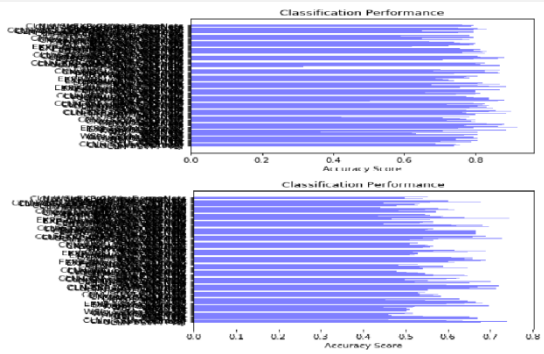

Fig 3 COMPARISON GRAPH 5,6Year Survival

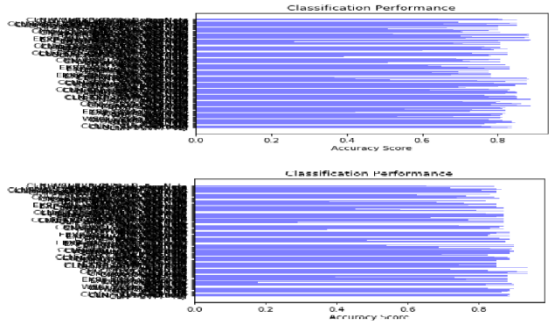

Fig 4 COMPARISON GRAPH 7,8 Year Survival

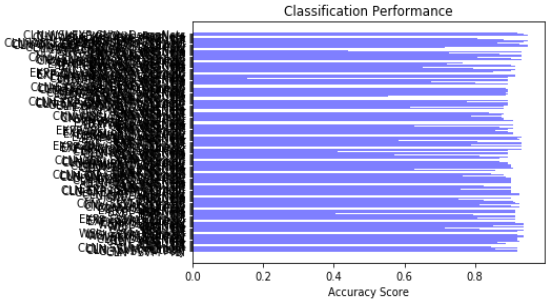

Fig 5 COMPARISON GRAPH 9 Year Survival

	ML Model	Accuracy	Precision	Recall	F1-Score
0	CLN - SVM - rbf	0.739	0.073	0.429	0.125
1	CLN - SVM - Poly	0.739	0.171	0.467	0.250
2	CLN - SVM - Utility	0.758	0.463	0.528	0.494
3	CLN - SVM - Linear	0.745	0.000	0.000	0.000
4	CLN - SVM - Sigmoid	0.689	0.049	0.154	0.074
...	...	...	...	...	...
175	CLN-WSI-EXP-CNV - NB	0.752	0.727	0.436	0.545
176	CLN-WSI-EXP-CNV- voting	0.863	0.455	0.789	0.577
177	CLN-WSI-EXP-CNV - VGGNets	0.789	0.182	0.462	0.261
178	CLN-WSI-EXP-CNV - ResNets	0.795	0.000	0.000	0.000
179	CLN-WSI-EXP-CNV - DenseNets	0.764	0.091	0.273	0.136

Fig 6 PERFORMANCE EVALUATION- Survival Prediction Models at 5-Year Survival

	ML Model	Accuracy	Precision	Recall	F1-Score
0	CLN - SVM - rbf	0.677	0.446	0.750	0.559
1	CLN - SVM - Poly	0.739	0.608	0.776	0.682
2	CLN - SVM - Utility	0.596	0.541	0.563	0.552
3	CLN - SVM - Linear	0.634	0.257	0.826	0.392
4	CLN - SVM - Sigmoid	0.460	0.324	0.393	0.356
...	...	...	...	...	...
175	CLN-WSI-EXP-CNV - NB	0.584	0.534	0.542	0.538
176	CLN-WSI-EXP-CNV- voting	0.615	0.479	0.593	0.530
177	CLN-WSI-EXP-CNV - VGGNets	0.540	0.096	0.467	0.159
178	CLN-WSI-EXP-CNV - ResNets	0.497	0.548	0.455	0.497
179	CLN-WSI-EXP-CNV - DenseNets	0.553	0.014	1.000	0.027

fig 7 PERFORMANCE EVALUATION- Survival Prediction Models at 6-Year Survival

	ML Model	Accuracy	Precision	Recall	F1-Score
0	CLN - SVM - rbf	0.839	1.000	0.839	0.912
1	CLN - SVM - Poly	0.826	0.985	0.836	0.905
2	CLN - SVM - Utility	0.783	0.889	0.857	0.873
3	CLN - SVM - Linear	0.839	1.000	0.839	0.912
4	CLN - SVM - Sigmoid	0.758	0.889	0.833	0.860
...	...	...	...	...	...
175	CLN-WSI-EXP-CNV - NB	0.696	0.723	0.900	0.802
176	CLN-WSI-EXP-CNV- voting	0.826	0.964	0.852	0.904
177	CLN-WSI-EXP-CNV - VGGNets	0.851	1.000	0.851	0.919
178	CLN-WSI-EXP-CNV - ResNets	0.814	0.956	0.845	0.897
179	CLN-WSI-EXP-CNV - DenseNets	0.801	0.942	0.843	0.890

Fig 8 PERFORMANCE EVALUATION- Survival Prediction Models at 7-Year Survival

	ML Model	Accuracy	Precision	Recall	F1-Score
0	CLN - SVM - rbf	0.888	1.000	0.888	0.941
1	CLN - SVM - Poly	0.888	1.000	0.888	0.941
2	CLN - SVM - Utility	0.882	0.944	0.925	0.934
3	CLN - SVM - Linear	0.888	1.000	0.888	0.941
4	CLN - SVM - Sigmoid	0.851	0.951	0.889	0.919
...	...	...	...	...	...
175	CLN-WSI-EXP-CNV - NB	0.720	0.788	0.871	0.828
176	CLN-WSI-EXP-CNV- voting	0.851	1.000	0.851	0.919
177	CLN-WSI-EXP-CNV - VGGNets	0.845	0.993	0.850	0.916
178	CLN-WSI-EXP-CNV - ResNets	0.652	0.730	0.840	0.781
179	CLN-WSI-EXP-CNV - DenseNets	0.839	0.985	0.849	0.912

Fig 9 PERFORMANCE EVALUATION- Survival Prediction Models at 8-Year Survival

	ML Model	Accuracy	Precision	Recall	F1-Score
0	CLN - SVM - rbf	0.919	1.000	0.919	0.958
1	CLN - SVM - Poly	0.919	1.000	0.919	0.958
2	CLN - SVM - Utility	0.857	0.899	0.943	0.920
3	CLN - SVM - Linear	0.919	1.000	0.919	0.958
4	CLN - SVM - Sigmoid	0.845	0.912	0.918	0.915
...	...	...	...	...	...
175	CLN-WSI-EXP-CNV - NB	0.807	0.837	0.955	0.892
176	CLN-WSI-EXP-CNV- voting	0.950	1.000	0.950	0.975
177	CLN-WSI-EXP-CNV - VGGNets	0.950	1.000	0.950	0.975
178	CLN-WSI-EXP-CNV - ResNets	0.938	0.987	0.950	0.968
179	CLN-WSI-EXP-CNV - DenseNets	0.919	0.967	0.949	0.958

Fig 10 PERFORMANCE EVALUATION- Survival Prediction Models at 9-Year Survival

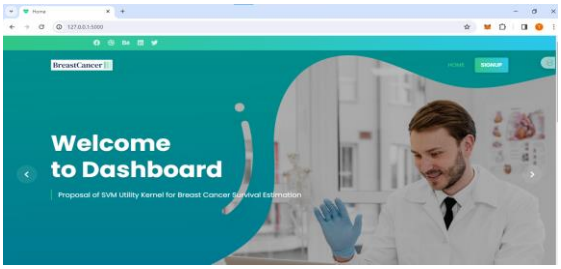

Fig 11 Home Page

Fig 12 Sign Up

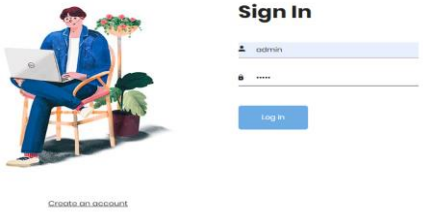

Fig 13 Sign In

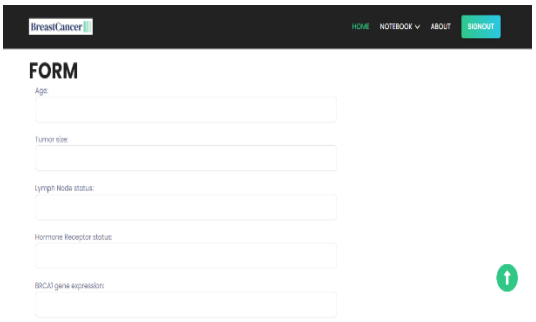

Fig 14 Upload Input Data

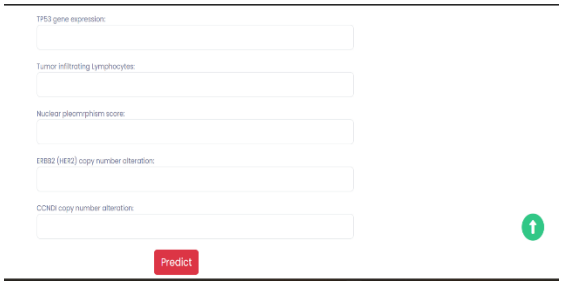

Fig 15 Upload Input Data

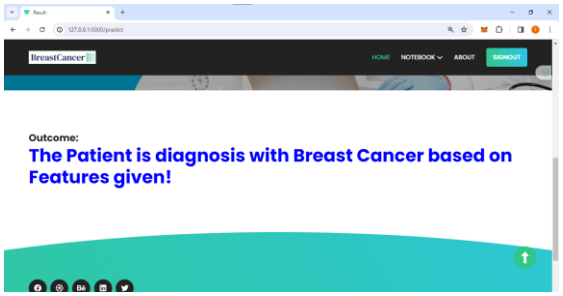

Fig 16 Predicted Result

5. CONCLUSION

In conclusion, the project highlights the importance of comprehensive and customized approaches in breast cancer survival prediction. By systematically assessing various machine learning algorithms, including tailored SVMs with utility kernels, traditional methods, and deep learning architectures like VGG-Nets and ResNets, the project demonstrates the significance of algorithmic customization for accurate predictions. Integration of diverse data sources, such as clinical data, whole-slide images, and molecular datasets, ensures a holistic understanding of breast cancer characteristics.

The extension efforts, particularly the exploration of a voting classifier, further enhance predictive capabilities, emphasizing the project's commitment to innovation in breast cancer prediction. The implementation of a user-friendly Flask-based deployment allows for easy access and early intervention, facilitating improved patient care.

6. FUTURE SCOPE

The feature scope for Study of utility kernal based svm for survival estimation of breast cancer encompasses a comprehensive set of clinical, imaging, and molecular features. This includes patient demographics such as age, gender, and ethnicity, tumor characteristics such as size, grade, and stage, treatment history, as well as molecular markers such as gene expression profiles and copy number variations. Additionally, imaging features extracted from whole-slide images contribute to the feature set, providing spatial information about tumor morphology and microenvironment.

REFERENCE

[1] G. M. Clark, “Do we really need prognostic factors for breast can cer?,” Breast Cancer Res. Treat., vol. 30, no. 2, pp. 117–126, 1994.
[2] L. R. Martin, S. L. Williams, K. B. Haskard, and M. R. Dimatteo, “The challenge of patient adherence,” Therapeutics Clin. Risk Man age., vol. 1, no. 3, pp. 189–199, Sep. 2005.
[3] C. Curtis et al., “The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups,” Nature, vol. 486, no. 7403, pp. 346–352, Apr. 2012.
[4] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, “Review the cancer genome atlas (TCGA): An immeasurable source of knowl edge,” Wspo»czesna Onkologia, vol. 1A, pp. 68–77, 2015.
[5] D. Delen, G. Walker, and A. Kadam, “Predicting breast cancer survivability: A comparison of three data mining methods,” Artif. Intell. Med., vol. 34, no. 2, pp. 113–127, Jun. 2005.
[6] K. Polyak, “Heterogeneity in breast cancer,” J. Clin. Investigation, vol. 121, no. 10, pp. 3786–3788, Oct. 2011.
[7] Z. Obermeyer and E. J. Emanuel, “Predicting the future– Big Data, machine learning, and clinical medicine,” New England J. Med., vol. 375, no. 13, pp. 1216–1219, Sep. 2016.