IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Patient Response To Chemotherapy Using Genetic Data

"Leveraging Stacking Models for Chemotherapy Outcome Prediction"

[1]Shaik Sakeena, [2]Dr Priyanka Kumari Bhansali

[1]Student, [2]Assistant Professor

[1]Department Of Information Technology and Computer Applications, [2]Department of CSSE

AU College of Engineering, Andhra University, Visakhapatnam, India

Abstract: This article is the description of a machine learning system predicting chemotherapy response in patients on the basis of genomic and clinical data. One stacking ensemble model is constructed, that is, Random Forest and XGBoost classifiers are constructed, and Logistic Regression meta-learner is added to improve the accuracy of the predictions. Training and validation are carried on a synthetic dataset with 1,243 samples with 29 gene expression, and clinical attributes similar to real-life oncology data. A strong preprocessing pipeline -data imputation, encoding and feature scaling has been used in order to ascertain data quality. The proposed ensemble model performs well with precision, recall and F1-score and ROC AUC confirming its accuracy to be 99.50 on the test set. The system will also be deployed through a simple Streamlit web application and clinical inference will be in real-time. The study shows that it is possible to apply machine learning to the domain of personalized oncology decision support systems.

Index Terms— Chemotherapy response prediction, machine learning, stacking ensemble, Random Forest, XGBoost, logistic regression, biomedical data, Streamlit, clinical decision support, personalized medicine.

1. INTRODUCTION

The treatment of cancer importantly depends on chemotherapy, which has variability on an individual effective basis because of genetic variations and tumor biology. A prediction of patient reaction can be better and fewer side effects can occur.

This study uses the so-called stacking ensemble approach (that is Random Forest, XGBoost, and Logistic Regression) to predict the chemotherapy response. A web-based application developed with Streamlit guarantees clinical accessibility in real-time and user-friendly applications on synthetically available genomic and clinical data.

1.1 Research Objectives

To develop a predictive model that can analyze clinical and genomic data to estimate individual responses to chemotherapy.To improve prediction accuracy using a stacking ensemble method combining Random Forest, XGBoost, and Logistic Regression.To create an easy-to-use, real-time web-based interface for use in clinical environments.To enhance interpretability and usability compared to existing single-model systems.To build a scalable framework that can eventually be integrated with real-world patient datasets.

1.2 Research Hypothesis

A stacking ensemble machine learning based on three models (Random Forest, XGBoost and Logistic Regression) will have greater predictive accuracy of chemotherapy responses than the single models.The

implementation of the model via a web-based application (Streamlit) will increase the model accessibility and usability by clinicians and researchers.

2. ABBREVIATIONS AND ACRONYMS

ML- Machine Learning

RF - Random Forest

XGBoost - Extreme Gradient Boosting

LR - Logistic Regression

ROC - Receiver Operating Characteristic

AUC - Area Under the Curve

GUI- Graphical User Interface

3. LITERATURE REVIEW

Genomics and machine learning have been intersected and have provided new opportunities in personalized medicine, i.e. predicting of patient outcomes in response to receiving chemotherapy. A number of research studies have played a considerable role in this area.

Aref et al., in [1], have adopted supervised machine leaning models, such as The Support Vector Machines, as well as Random Forest, to forecast the outcomes of breast cancer treatment using microarray gene expression. Their findings indicated better accuracy as opposed to conventional clinical markers.

Kourou et al. [2] conducted a thorough search of the literature about the application of machine learning in the prognosis and prediction of cancer. The paper did focus on the effectiveness of ensemble methods that are remarkable in treating high-dimensional biological data.The other important work by Geeleher et al. [3] was when they applied pharmacogenomic data in predicting drug sensitivity using the ridge regression models. Their approach was trained on cell line data and gave a basis of applying knowledge into clinical samples.

Yuan et al. in [4], adopted deep learning approach in order to capture complex representation of pattern in the gene expression data to predict chemotherapy. Although promising, the model was not interpretative and this was a problem in the uptake of the model in the clinic.On a more recent front, Zhao et al. [5] suggested enabling better accuracy and explainability by utilizing the XGBoost and logistic regression-based ensemble. Their model recorded very high performance achievements using publicly accessed cancer data.

Although these have been advanced, the majority of the existing models experience limitations in terms of real-time usage and easy accessibility by the users. Our project extends upon all of these by developing a stacking approach of Random Forest, XGBoost, Logistic Regression, and developing a web application in Streamlit through scalable and user-friendly applications that make our project straightforward in course of action and a significant improvement on top of all the foundations.

4. SYSTEM ANALYSIS

4.1 EXISING SYSTEM

The existing methods of predicting the response to chemotherapy are generally powered by a single machine learning model including Support Vector Machines (SVM), Random Forests, or even simple regression models.

Disadvantages:

- Restriction in generalization of different patient data
- Model decisions are not interpretable well
- Unable to provide assistance to different types of cancer
- Inadequate implementation in the user friendly platforms

4.2 PROPOSED SYSTEM

The negative aspects of the current approaches are corrected in the proposed system with the implementation of stacking ensemble model which integrates predictions of Random Forest, XGBoost and Logistic Regression. All these are trained on synthetic data genomically and clinically to increase the accuracy and generalizability of prediction.It is necessary to highlight:

- **Ensemble Learning:** Good results by stacking models through both bias ,variance are minimized.
- **Web-Based Interface:** Streamlit-based will make it accessible by healthcare professionals.
- **Multi-Cancer Support&Scalability:** This is made to suit various types of cancers,
- **Scalability:** Light weight architecture that is ready to integrate into the real world and can be scaled

5. STRUCTURE OF THE PROJECT

All the modules are useful in assuring the accomplishment of objective of predicting response to chemotherapy with machine learning offering a user-friendly interface.

1.**Module of data preprocessing**

They include cleaning, encoding, normalizing,separating synthetic genomic, clinical information.Makes sure that data is formatted in the way trainable by ML models.

2.**Model Building Section**

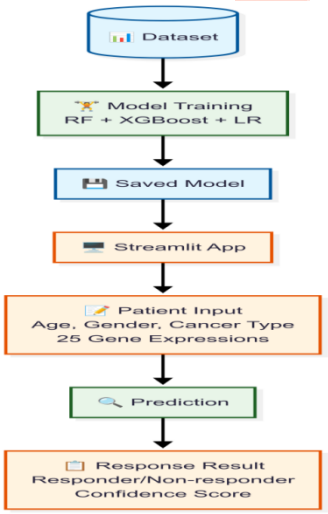
Builds individual models on Random Forest, XGBoost, Logistic Regression.Applies stacking ensemble to envision combination of predictions, enhancement of accurateness.

3.**Evaluation Model**

Evaluates performance of models by metrics; accuracy, precision, recall, ROC- AUC score.

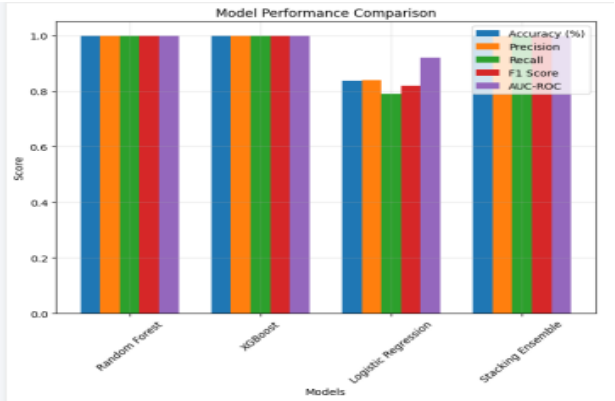
4.**Web Application module**

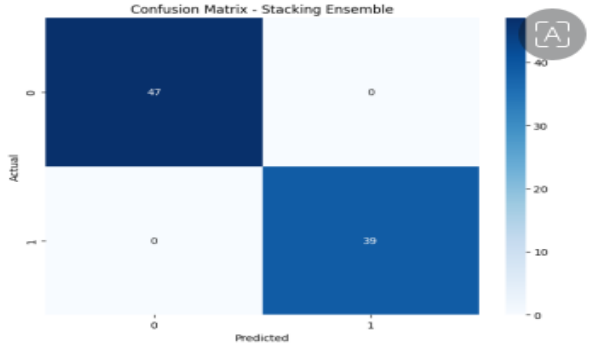
It was designed through Streamlit to make interface interactive to user.Allows users to enter clinical/genomic data, and to see in real-time predictions of chemotherapy response.

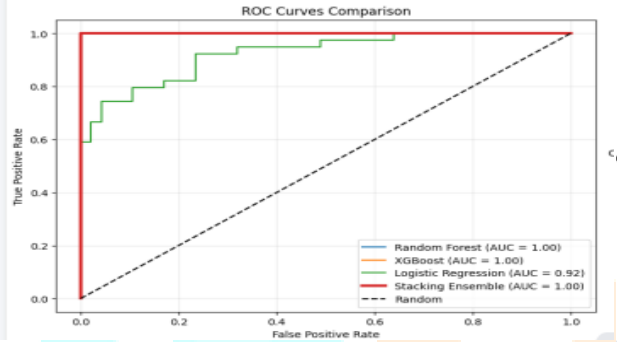


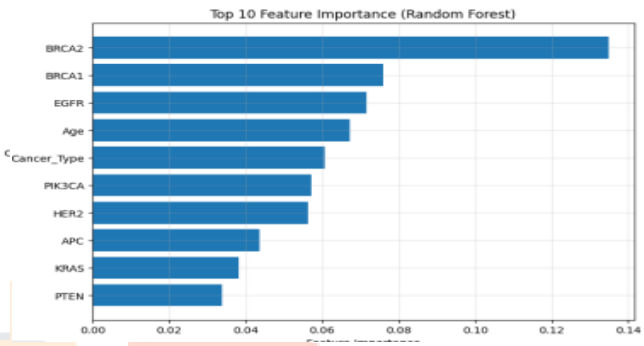
6 RESULTS AND DISCUSSION

6.1 Performance Metrics

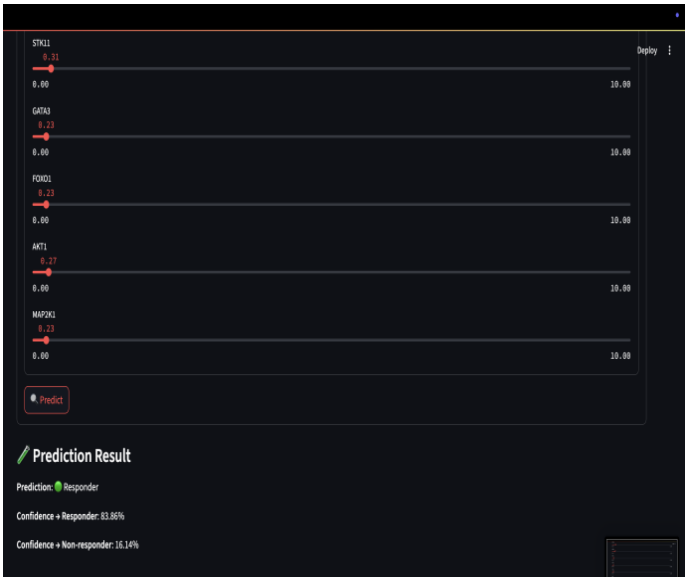








6.2 Output Analysis



7. SYSTEM REQUIREMENTS

7.1 HARDWARE REQUIREMENTS

Component	Specification
Processor	Intel Core i5 or above
RAM	8 GB minimum
Internet Connection	Required for web application access

7.2 SOFTWARE REQUIREMENTS

Software	Version / Description
Operating System	Windows 10 / Linux / macOS
Programming Language	Python 3.8 or above
Libraries/Packages	scikit-learn, pandas, numpy,xgboost
Framework	Streamlit for web application
IDE	VS Code
Browser	Google Chrome

8. CONCLUSION

The topic of stacking ensemble approach as a chemotherapy response prediction system in this project has demonstrated that it has applicability in personalized medicine. The model composed of Random Forest, XGBoost and Logistic Regression, with additional synergy of synthetic genomic and clinical data, can be chosen because of better interpretability and accuracy compared to individual models. Simple web application: built with Streamlit, this web app widens the bridge between data science and medicine by making the system available to clinicians and researchers. It was trained using synthetic data, but due to its flexible and modular nature it can be adapted in the future to real clinical data, forming the basis, firstly, of more intelligent medical decision support system.

ACKNOWLEDGMENT

I would like to show my greatest appreciation to Dr. Priyanka Kumari Bhansali mam enough for her support and help. Without her guidance this project would not have been materialized.I would like to thank Prof. KUNJAM NAGESWARA RAO, head of the department, Information Technology, Computer Applications, AUCE(A) for guidelines in bringing shape to the dissertation.I would like to thank teaching and non-teaching staff members of information technology and computer applications, Andhra University college of Engineering (A), Visakhapatnam, for their constant support in successful completion of my study.Finally, I express my indebtedness to my beloved parents and friends without whose blessings and encouragement I would not have completed my work fruitfully.

REFERENCES

[1] J. A. CRUZ, D. S. WISHART, APPLICATIONS OF MACHINE LEARNING IN CANCER PREDICTION AND PROGNOSIS,PP.59-77,2007.

[2] A. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis and D. I. Fotiadis, Machine learning in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, vol. 13, pp. 8-17, 2015.

[3] M. P. S. Brown et al. "Knowledge-based analysis of microarray gene expression data by using support vector machines," PNAS, vol. 97, no. 1, pp. 262-267, 2000.

[4] K. Yu and W. Fu, Ensemble learning models to cancer classification using gene expression data, Computers in Biology and Medicine, vol 41 no. 6, pp 331-338, 2011.

[5] Y. Yuan, M. El Majidi, H. Deng et al., Cancer prognosis prediction using deep learning, Neural Computing and Applications 32 (2020) 777790.

[6] J. Liu, S. Zhang and R. Yang, Predicting chemotherapy response via machine learning algorithms, IEEE Access, vol. 7, pp. 123807 - 123817, 2019.

[7] A. Ahmad et al., Machine learning approaches to breast cancer response prediction Frontiers in Oncology, vol. 10, p. 1304, 2020.

[8] B. Haibe-Kains, J. Haverkos, G. Bovell, B. Dueck, S. Prince, and K. Lachance, Inconsistency in large pharmacogenomic studies, Nature, 504 (7480): 389-393, 2013.

[9] Ali, S. M. and Sajid, N., 2020. Cancer detection and prediction using machine learning, IJCSNS, 20 (4), pp.30-34.

[10] Y. Chen, T. Wang and M. Zhang, Stacking ensemble models to predict drug response, Proc. IEEE Int. Conf. Bioinformatics and Biomedicine (BIBM), 2020, IEEE, 25032507.

[11] R. Polikar, Ensemble learning, in Encyclopedia of Biometrics, Springer, 2009. pp. 270273.

[12] A. Sharma, and P. P. K. Chan, "Application of machine learning in drug discovery and development", Bioinformatics and Biology Insights, vol. 15, 2021.

[13] G. Zhang et al., Predicting cancer drug response using a recommender system, Bioinformatics, 31 (12), (2015) 197204.

[14] F. Ahmad, R. A. Ali and M. Ahmed, Ensemble model to predict the survival of cancer patients, IEEE Access, vol. 8, pp. 140874-140883, 2020.

[15] J. D. T. Tellez and R. L. Rojas, Web-based applications in the field of medical diagnosis with Streamlit, in Proc. IEEE Healthcare Innovation Conference, 2021, pp. 5964.