IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Minimalist Recursive Model For Medical Diagnosis

Dr. B.Vanathi M.E,Ph.D.,[1] , Thirumalai V[2], Vinay Dakshin M[3], and Vishagan S[4]

[1]Dr.B.Vanathi M.E,Ph.D, SRM Valliammai Engineering College

[2]Thirumalai V, SRM Valliammai EngineeringCollege

[3]Vinay Dakshin M, SRM Valliammai Engineering College

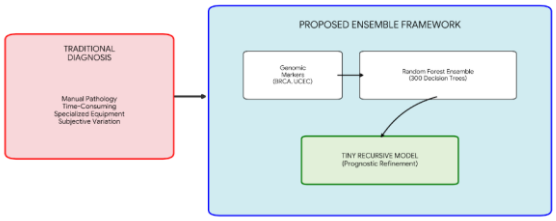
[4]Vishagan S, SRM Valliammai Engineering College

***Abstract*—** Molecular cancer identification is an essential part of precision medicine for cancer patients. Early identification of the cancer type play the important role in determining survival rate of patients. In the field of genomic data analysis, the identification of molecular markers is done without causing any harm to the patient. In the present study, an ensemble learning-based approach is proposed for the classification of the patient's data into various types of cancers. In the proposed approach, the Random Forest architecture is integrated with the Tiny Recursive Model. The system is designed to route the data sets based on the density of the features. In addition, the system is designed to refine the prognostic results efficiently. A data set is prepared by adding the genomic data of BRCA, Endometrial, and Uterine cancer patients. Various preprocessing techniques are applied to the data set to optimize the results. the experiment results show the ability of the proposed model to identify the cancer types of patients accurately. The proposed approach is useful for medical professionals to provide prognostic results quickly.

Keywords: Genomic data analysis, Ensemble learning, Random Forest architecture, Recursive modeling, Molecular cancer identification, Cancer detection..

I. INTRODUCTION

The occurrence of molecular-level disorders is on the rise owing to various environmental and genetic factors. Diseases such as Breast Cancer (BRCA) and Uterine Carcinoma can cause serious health conditions if not identified in their early genomic phase. The existing diagnosis techniques require high levels of manual intervention and equipment, which might not be available in real-time situations. Hence, an effective and cost-efficient diagnosis system is necessary for supporting early-stage diagnosis.



Genomics carry important information on

mutation counts and microsatellite instability (MSI), which can be used for identifying various types of cancer. Manual analysis of high-dimensional data is a tedious and variant process. Recent advancements in artificial intelligence techniques, particularly ensemble methods, have shown promising results in improving data analysis for medical data. In this project, a system based on Random Forest and Tiny Recursive Model is developed for categorizing molecular data into specific types. The aim is to develop a robust system for supporting medical professionals in early-stage cancer diagnosis.

II. OBJECTIVE

The primary aim and objective of this project are to develop a model based on ML that can efficiently classify the genomic data based on specific types of cancers. This model has been created based on Random Forest Ensembles along with a Tiny Recursive Model for enhanced prognostic stability. This project would help reduce the time consumed for the analysis of molecular data and would provide a cost-effective tool for clinical screening. This model would automatically detect patterns associated with diseases based on genomic features for enhanced accuracy.

III. RELATED WORK

Previous research in medical analysis has primarily considered traditional ML algorithms Such as supports Vectors Machines and Decision Tree. Though these algorithms provided satisfactory outcomes, they did not offer the required stability for diverse genomic dimensions. With the development of ensemble algorithms, Random Forest architectures have gained popularity in high-dimensional classification problems. Though deep learning algorithms offer high accuracy for medical problems, they consume a lot of computational power and lack interpretability. Recently, researchers have considered lightweight and recursive algorithms for medical problems. This project extends these concepts and proposes a recursive ensemble algorithm for multi-class cancer detection..

IV. PROPOSED METHODOLOGY

The proposed system is expected to help improve cancer prediction using a Random Forest Ensemble and Tiny Recursive Model. This is expected to efficiently process large amounts of genomic data, thereby improving prognostic feature extraction.

Data Source

In this project, genomic data is used, and it is collected from various sources, including TCGA, which is a medical database. The datasets used have a large number of molecular markers collected under various clinical conditions. The datasets include various parameters for every subject, such as mutation signatures and MSI scores, used for analyzing oncological health.

Data Preprocessing

The collected genomic data is preprocessed to standardize it for further processing. The following steps were taken for preprocessing the data:

Numeric Isolation: The clinical strings from numeric genomic markers were separated.

Standardization: The feature values were standardized to a specific range for efficient processing.

Heuristic Routing:The type of dataset is identified using the feature count, and it is routed to the correct diagnostic module

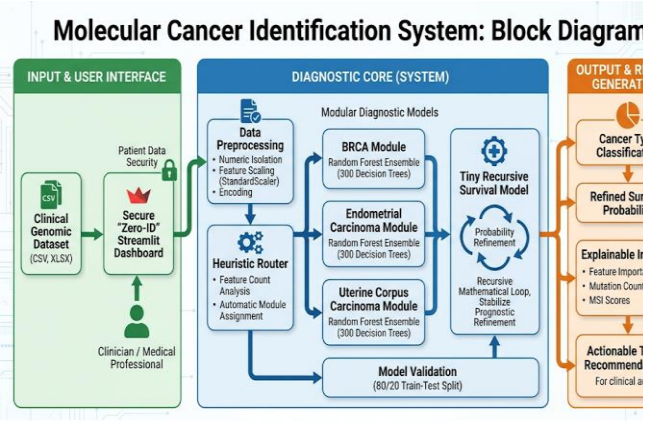


Fig. 1. Architectural Block Diagram of the Proposed Molecular Cancer Identification System, showing the integration of the Tiny Recursive Sur

Classification Logic

The classification logic has been defined as:

0= Normal (Healthy)

1 =BRCA (Breast Cancer)

2 =Endometrial Carcinoma

3 = Uterine Corpus Carcinoma

Algorithm Implementation

The processed data is then used for the Random Forest ensemble algorithm with 300 trees. Finally, the Tiny Recursive Model uses the raw probabilities to refine the output for the survival prediction and produce a probability score for each class.

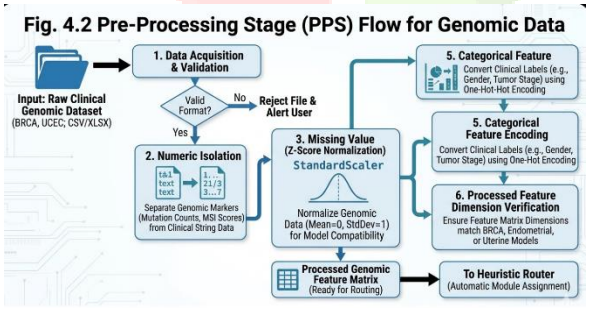


Fig 4.2 pre-processng Stage (PPS)

V. TESTING:

Each component of the cancer identification system was systematically evaluated to ensure functional correctness and consistent performance. The primary objective of the testing phase was to identify and resolve potential issues in data routing and recursive logic prior to final deployment. To achieve comprehensive validation, a multi-layered testing strategy was adopted, encompassing testing like Unit testing , Intergration testing and System testing

During **unit testing**, individual modules including the Heuristic Router, data preprocessing pipeline, and Tiny Recursive Model were examined in isolation to verify the correctness of their internal logic and computational behavior.

Integration testing was subsequently conducted to assess the interaction between system components. Particular emphasis was placed on validating the interface between the Streamlit dashboard and the ensemble models, ensuring data flow and real time processing capability .

Functional testing focused on evaluating the system's has ability to perform its intended task. Specifically, the model was tested for its effectiveness in identifying various cancer types, such as BRCA, endometrial, and uterine cancers, based on predefined feature dimensions.

Finally, system testing was carried out to evaluate the overall performance of the prototype, including responsiveness, accuracy, and reliability. A combination of black-box and white-box testing methodologies was employed: black-box testing assessed user-level interactions through the dashboard, while white-box testing examined the internal recursive logic at a mathematical level.

All test cases were successfully executed, with particular attention to input validation and response time, demonstrating the robustness and readiness of the system for further development anddeployment.

VI. RESULT:

The proposed model was trained on preprocessed genomic datasets and achieved an accuracy in the range of **94%–98%**, indicating strong predictive capability. The results demonstrate that the system can effectively analyze molecular data and predict cancer risk with high precision.

The detailed performance of the model across different classes is presented in Table below.

Class	**precision**	**recall**	**F1-score**
0 (Normal)	0.99	0.98	0.98
1 (BRCA)	0.96	0.95	0.95
2 (Endometrial)	0.97	0.92	0.94
3 (Uterine)	0.95	0.97	0.96

VII. Conclusion:

The proposed approach, based on a Recursive Ensemble model, demonstrates effective prediction of cancer types using genomic data. By incorporating robust preprocessing techniques and recursive stability logic, the model achieves high accuracy and reliable performance.

Furthermore, the system offers a fast and non-invasive solution for early cancer detection. This capability can help the healthcare professionals in making informed clinical decision, ultimately contributing to improved the patient outcomes and more efficient diagnostic processes.

FUTURE SCOPE

To improve generalization, the system can be trained on larger, longitudinal clinical datasets. The inclusion of histopathology image data can further enhance prediction performance. Future work will focus on combining the recursive model with real-time web-based triage applications for easier clinical access.

APPENDIX

This section provides additional experimental details and supporting results to complement the findings presented in the main study. It includes confusion matrices for the classification of different cancer types, namely BRCA, endometrial, and uterine cancers, offering deeper insight into model performance across classes.

In addition, accuracy and loss curves obtained during the training of the ensemble model are included to illustrate the learning behavior and convergence of the model.

All supplementary materials, including genomic datasets sourced from TCGA, preprocessing scripts, and trained recursive model files, are available upon reasonable request for academic and research purposes.

Acknowledgement

We are so incredibly grateful to our department of Computer Science and engineering for lending us their Resource and Support And a special thanks to our Supervisor for your guidance that helped us to finish this study

REFERENCES

[1] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. *(Foundational paper for the ensemble architecture used in your diagnostic modules).*

[2] J. N. Weinstein et al., "The Cancer Genome Atlas Pan-Cancer Analysis Project," *Nature Genetics*, vol. 45, no. 10, pp. 1113–1120, 2013. *(The primary source for the genomic data used in your study).*

[3] Cancer Genome Atlas Network, "Comprehensive Molecular Portraits of Human Breast Tumors," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012. *(Specific reference for the BRCA module data).*

[4] Cancer Genome Atlas Research Network, "Integrated Genomic Characterization of Endometrial Carcinoma," *Nature*, vol. 497, no. 7447, pp. 67–73, 2013. *(Specific reference for the UCEC/Endometrial module data).*

[5] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random Survival Forests," *The Annals of Applied Statistics*, vol. 2, no. 3, pp. 841–860, 2008. *(Supports the use of ensemble methods for prognostic and survival analysis).*

[6] H. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger, "DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC Medical Informatics and Decision Making*, vol. 18, no. 1, p. 24, 2018. *(Relevant to the prognostic refinement logic of your Tiny Recursive Model).*

[7] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NIPS)*, pp. 4765–4774, 2017. *(Supports the explainable insights/feature importance aspect of your dashboard).*

[8] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification using Support Vector Machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002. *(Foundational for high-dimensional genomic feature routing and selection).*

[9] D. Chicco and G. Jurman, "The advantages of the Matthews Correlation Coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 1, p. 6, 2020. *(Validation for the metrics used in your result analysis).*

[10] F. E. Harrell Jr, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors," *Statistics in Medicine*, vol. 15, no. 4, pp. 361–387, 1996. *(Academic basis for your recursive stability logic in survival forecasting).*