**Holistic Assessment of malignant epithelial ovarian tumours based on clinical data and biomarkers**

*Project Report Submitted*

*to*

**MANIPAL ACADEMY OF HIGHER EDUCATION**

*For Partial Fulfilment of the Requirement for the*

*Award of the Degree Of*

**Bachelor of Technology**

*in*

**Computer and Communication Engineering**

*By*

RAGHVENDRA MISRA

Reg. No.– 210953106

KHUSHI AGARWAL

Reg. No.– 210953108

*Under the guidance of*

Mr. Chetan Sharma Dr. Manjula Shenoy K

Assistant Professor-Senior Scale Professor

Department of I&CT Department of I&CT

Manipal Institute of Technology Manipal Institute of Technology

Manipal, Karnataka, India Manipal, Karnataka, India



**INDEX:**

|  |  |
| --- | --- |
| PAGE NUMBER | CHAPTER |
| 1 | FRONT PAGE |
| 2 | INDEX |
| 3 | INTRODUCTION |
| 4 | LITERATURE SURVEY |
| 9 | METHODOLOGY |
| 15 | RESULTS AND DISCUSSION |
| 19 | CONCLUSION |
| 21 | REFERENCES |

**INTRODUCTION**

The seventh most prevalent illness in women is ovarian cancer (OC), which carries a high lifetime risk. Low survival rates are caused by late-stage diagnosis and a lack of early indications, which highlights the critical need for efficient diagnostic instruments. Our goal is to provide comprehensive research of malignant epithelial ovarian tumours by integrating clinical data and biomarkers.

Crucial Biomarkers

-> CA125

-> HE4

-> NEU

These biomarkers represent the cutting edge of our research, each providing a distinct window into the complex molecular and clinical features of ovarian cancers. In addition to the well-known ovarian cancer marker CA125, HE4 and NEU complete a trio of biomarkers that together have the potential to improve the sensitivity and specificity of our diagnostic models.

The following primary goals are the focus of this investigation:

Early Detection: Using clinical data and biomarkers together, developing new approaches for the early detection of malignant epithelial ovarian cancers.

Biomarker importance: Identifying patterns suggestive of ovarian cancers by revealing the importance and interaction of CA125, HE4, and NEU with a wide range of clinical characteristics.

Comprehensive Analysis: To conduct rigorous and statistically sound analytical studies, advanced machine learning models are applied to a broad dataset that includes blood samples, general chemistry medical tests, and ovarian cancer markers.

Keywords: malignant epithelial ovarian tumours, biomarkers, clinical data, machine learning, gene sequencing, early detection, prognostication.

**LITERATURE SURVEY**

**PAPER 1**

Debaditya Chakroborty et.al [[1]](https://doi.org/10.1101/2023.07.24.550346)

By utilizing eXplainable AI (XAI), researchers sought to find genetic markers for targeted cancer therapy with a particular emphasis on high grade serous carcinoma (HGSC). Their approach combined SHapely Additive Explanations (SHAP) and XGBoost to identify pertinent biomarkers, resulting in clear predictions and economical hypothesis testing. Although the study has been successful in predicting the probability of a 5-year survival, it recognizes that meaningful confirmation requires human trials, particularly considering the reliance on mouse tumor models. Although the method provides biological insights, it is still difficult to prove cause-and-effect linkages because some biomarkers show context dependency. This study emphasizes the promise of XAI and probabilistic methods in cancer research, but it also emphasizes the need for additional experimental validation.

**PAPER 2**

Kristofer Linton-Reid et.al [[2]](https://doi.org/10.1101/2023.04.26.23289155)

An end-to-end integrated method is used by researchers under the direction of Kristofer Linton-Reid to tackle the problem of risk classification in high-grade serous ovarian cancer (HGSC). This includes using machine learning (ML) and deep learning (DL) models, such as CNNs and Cox proportional hazard models, as well as radiomics analysis for non-invasive tumor information extraction and CT scan segmentation to locate ovarian masses. The intricacy of DL modelling and the non-invasiveness of radiomics analysis are among the benefits. DL and ML models are resource-intensive, and big medical picture datasets are required for radiomics feature validation. These are some of the drawbacks. Consistently exhibiting higher prediction accuracy, the Permutation-Variable Importance Random Forest - Random Survival Forest (PVIRF-RSF) model offers a more accurate data-driven method and shows promise in risk assessment.

**PAPER 3**

Haoxin Zhang et.al [[3]](https://doi.org/10.1016/j.compbiomed.2022.106432)

A novel random forest-based metabolic risk model for ovarian cancer prognostic assessment and identification of metabolism-related therapeutic targets is presented by Haoxin Zhang and colleagues. Patients are categorized into low and high metabolic subtypes based on the model's identification of 17 metabolic pathways associated with prognosis. Several datasets are used in the study to increase sample size, and different biological studies are used to identify subtype differences. WGCNA is utilized to identify genes linked to subtypes, and an XGBoost classifier is employed to predict subtypes. Although the method offers insightful suggestions for novel pharmacological targets and customizable strategies, its drawbacks include a lack of clinical validation, possible bias in the TCGA data, and the requirement to evaluate generalizability across separate datasets. When predicting metabolic conditions, the XGBoost Classifier shows excellent accuracy and a high area under the ROC curve, enhancing the understanding of cellular metabolisms role in ovarian cancer.

**PAPER 4**

Mingyang Lu et.al [[4]](https://doi.org/10.1016/j.ijmedinf.2020.104195)

Using ML approaches and the MRMR feature selection method on a dataset consisting of 49 variables, Mingyang Lu and colleagues address the prediction of ovarian cancer through machine learning. A decision tree model performs better than techniques like logistic regression and risk of ovarian malignancy algorithm(ROMA) when it is trained on one dataset and tested on another that includes both actual and control instances. The method's benefits include better performance over ROMA, a large dataset providing a comprehensive view, and reliance on two biomarkers (HE4 and CEA) for clinical simplicity. The dataset's exclusive Chinese patient population, possible bias resulting from missing data imputing, and the requirement for additional clinical validation and benchmarking are its drawbacks. Even though the study's early ovarian cancer detection method beat ROMA, it still highlights the need for additional validation, benchmarking, and generalizability across various populations.

**PAPER 5**

Zhong Yu et.al [[5]](https://journals.lww.com/md-journal/Fulltext/2022/09090/Identification_of_prognosis_related_hub_genes_of.59.aspx?context=LatestArticles)

Through a multi-step process that includes building protein-protein interaction networks, weighted gene co-expression analysis (WGCNA), survival analysis, qRT-PCR gene expression validation, mutation analysis, functional pathway analysis, drug sensitivity analysis, and assessment of tumor mutation burden (TMB), Zhong Yu and Ouyang Ling set out to identify prognosis-related hub genes in ovarian cancer. Comprehensive datasets include clinical data and the discovery of novel biomarkers (ALDH1A2, CLDN4, GPR37) are two benefits of the study. Analysis of functional pathways illuminates the mechanisms underlying ovarian cancer. Nevertheless, qRT-PCR's limited sample size, inconsistent findings with earlier research, and the critical requirement for additional experimental validation are some of its drawbacks. Although the study identifies three putative prognostic hub genes that provide enhanced understanding of ovarian cancer, it emphasizes the need for larger datasets and experimental validation to fully understand the implications and roles of these genes.

**PAPER 6**

Ishleen Kaur et.al [6] – NIH (National Library of Medicine)

A comprehensive method for predicting cancer survival is presented by Ishleen Kaur and colleagues in MATLAB. It includes data collection, preprocessing, and classification using SVM, k-NN, and decision trees. Clinical features that provide a comprehensive picture of the patient's health are taken into account, such as FIGO substage, ascites presence, and CA-125 levels. The results can be applied in clinical settings, and the use of sequence mining techniques for treatment sequences adds a dynamic element. However, the study's specificity to ovarian cancer limits generalizability, and ethical restrictions restrict dataset sharing, which hinders validation. With 2-month intervals between treatments, the integrated approach with ensemble techniques with sequential pattern missing achieves a noteworthy 76.4% accuracy. Although the study shows promise for improving prognostic accuracy, it also highlights the need for additional validation and acknowledges the insightful information it offers about survival prediction for advanced ovarian cancer patients.

**PAPER 7**

Eleni Karamouza et.al [7] - MDPI

Eleni Karamouza and colleagues use patient data from GCIG meta-analysis studies to investigate the predictive potential of CA-125 in newly diagnosed advanced ovarian cancer, adhering to PRISMA-IPD and TRIPOD guidelines. The study optimizes CA-125 monitoring by focusing on overall survival and the prognostic value of CA-125 levels at 3 and 6 months, using a large dataset of 5573 patients. The nonlinear nature of CA-125 over time is addressed by hierarchical linear mixed models. The study recognizes that CA-125 is susceptible to biological variation and measurement error, despite its large sample size and potential benefits in improving CA-125 monitoring. Logistically, it could be difficult to continue monitoring for longer than three months. The results of the study demonstrate that CA-125 measurements at three months can reliably forecast overall survival for up to 48 months, achieving an AUC of .75 at 24,36 months and .74,.75 at 48 months. This approach is useful and applicable to support patient treatment and enhance CA-125 monitoring.

**PAPER 8**

Weitong huang et.al [8]

Using ovarian cancer as a case study and machine learning (ML) and Shapley analysis, Weitong Huang and colleagues present an approach to explainable discovery of disease biomarkers. The study provides additional insights by improving the readability of ML algorithm decisions for subject-matter experts through the integration of ML and explainable AI (XAI) techniques. XAI techniques solve accountability issues by providing transparent model predictions, which is essential in sensitive medical applications. However, the completeness and quality of the data are crucial for the effectiveness of ML and XAI algorithms, which can sometimes struggle with noise or sparsity. It takes domain knowledge to choose ML algorithms and analytical methods correctly. The case study highlights the benefits of the ML pipeline over conventional statistical techniques as well as its efficacy and consistency. The resultant guidelines provide a useful foundation for implementing XAI in research, informing clinicians and explaining and validating cancer biomarkers

**PAPER 9**

Lili Fan et.al [9]

Using transcriptome-wide data mining, Lili Fan and colleagues address the identification of a gene set correlated with immune status in ovarian cancer. Tumor Mutation Burden (TMB) and immune-related gene expression profiles are two important biomarkers that are evaluated in this study using data mining techniques. TMB scores are computed, and single-sample Gene Set Enrichment Analysis (ssGSEA) is performed for immune gene sets. Although integrating TIME and TMB data to assess the immune status of ovarian cancer comprehensively improves patient outcome predictions, doing so may be limited by the specialized knowledge and resources required for the integration of diverse data sources and the ensuing bioinformatics analysis. A set of genes, including CXCL13, FCRLA, PLA2G2D, and MS4A1, has been found to be substantially correlated with a better prognosis in cases of ovarian cancer. The study offers a potential solution to challenges in immunotherapy for ovarian cancer by emphasizing the significance of identifying gene sets linked to immune status in ovarian cancer.

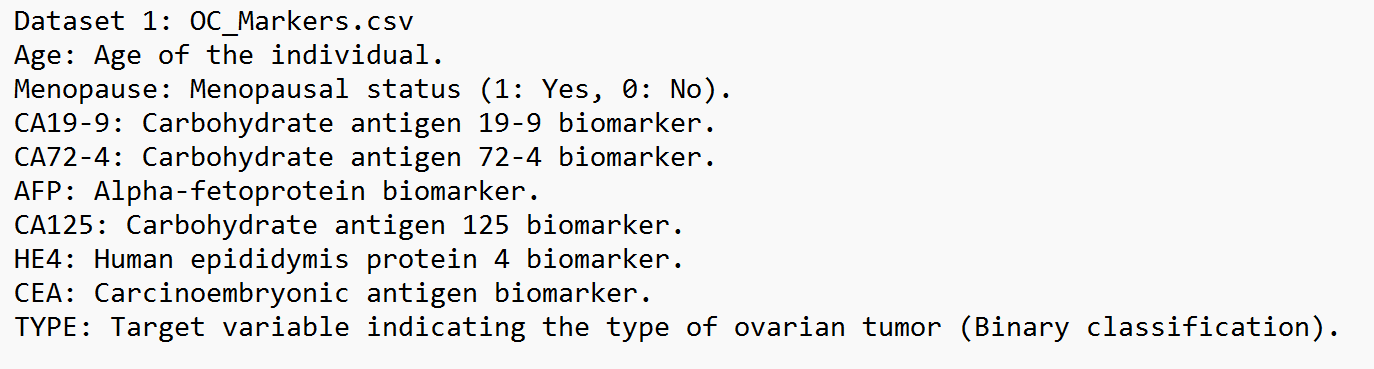
**PAPER 10**

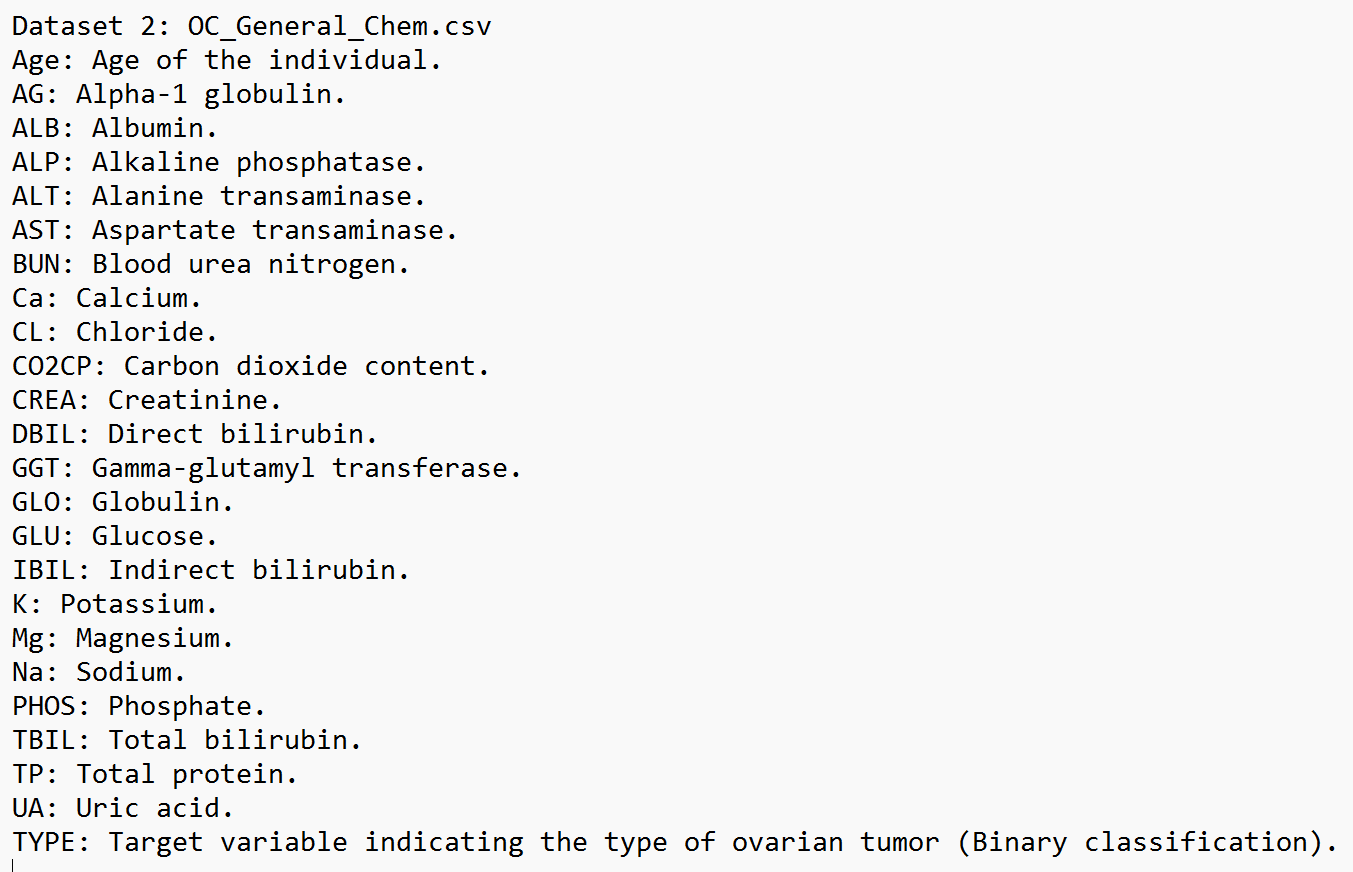
Lingyan Yuan et.al [10]

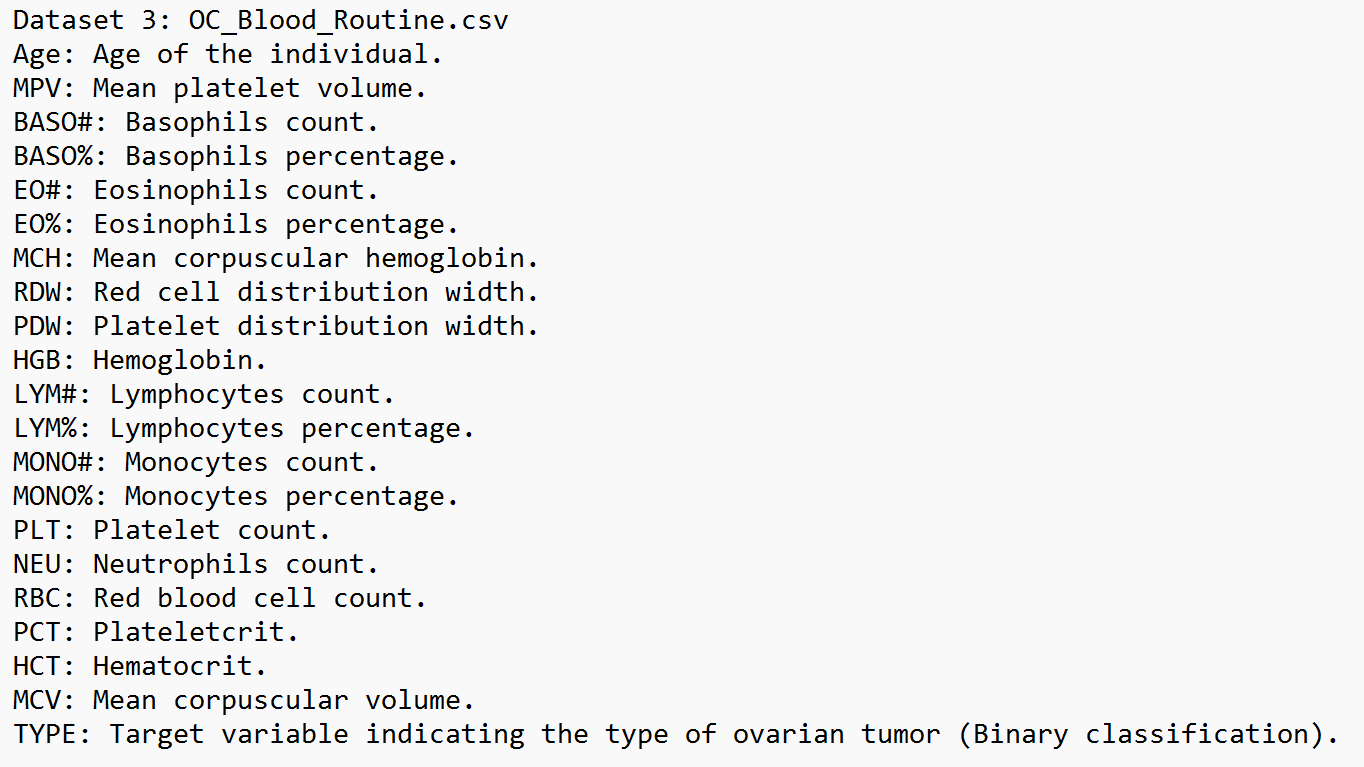
Taking heterogeneity into account, Lingyan Yuan and colleagues use integrative data mining and meta-analysis to examine the prognostic significance of the microRNA-200 family in a range of human malignant neoplasms. To evaluate the relationship between miR-200 family expression and patient prognosis, clinical data from the identified studies are combined, and hazard ratios with 95% confidence intervals are computed. The Cancer Genome Atlas is the source of miRNA expression data for members of the miR-200 family (TCGA). The review's benefit is that it evaluates the prognostic significance of miR-200 family expression heterogeneity by integrating clinical and molecular data from TCGA. However, variability may be introduced and results of meta-analyses may be impacted by heterogeneity in clinical data and miR-200 family expression across studies. With 4,644 patients and 15 different tumor types included in the meta-analysis, the miR-200 family is recognized as an independent protective factor for various tumours.

**METHODOLOGY**

We have taken 3 different datasets of 350 women and integrated it into one such that it has a wide variety of attributes to analyse the data and make appropriate predictions. The attributes of datasets are as follows:

****

****

****

We have also taken another data set that contains Gene Expression RNA seq data for 800 patients which can help us determine the type of cancer. This dataset is only used to apply the Artificial Neural Network model.

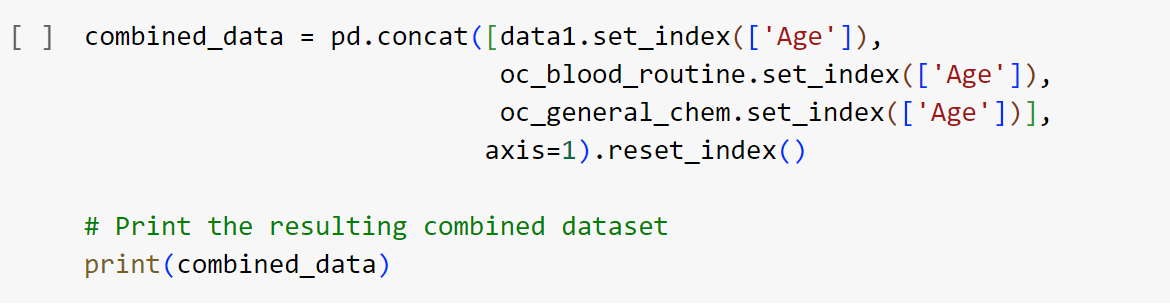
There are 8000 attributes, each one is given a dummy name gene\_XX for ease of understanding and mapping. The last column helps in identifying the type of cancer.

Dataset name: cancer\_gene\_expression.csv

**DATA PRE-PROCESSING:**

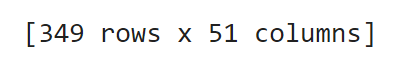
Data preparation is a crucial first step in preparing raw datasets for analysis and model construction. It employs multiple techniques to correct missing values, enhance the quality of the data, and ensure that machine learning algorithms can use it. The following are some essential steps in data preprocessing:

* Data Integration
* Handling Missing Data
* Dealing with Duplicates
* Handling Outliers
* Data Scaling

****

This is how we have integrated 3 datasets, namely OC\_Marker, OC\_General\_Chem and OC\_Blood\_Routine into 1 combined dataset and used this for the data analysis.

There are no NULL values in our final dataset.

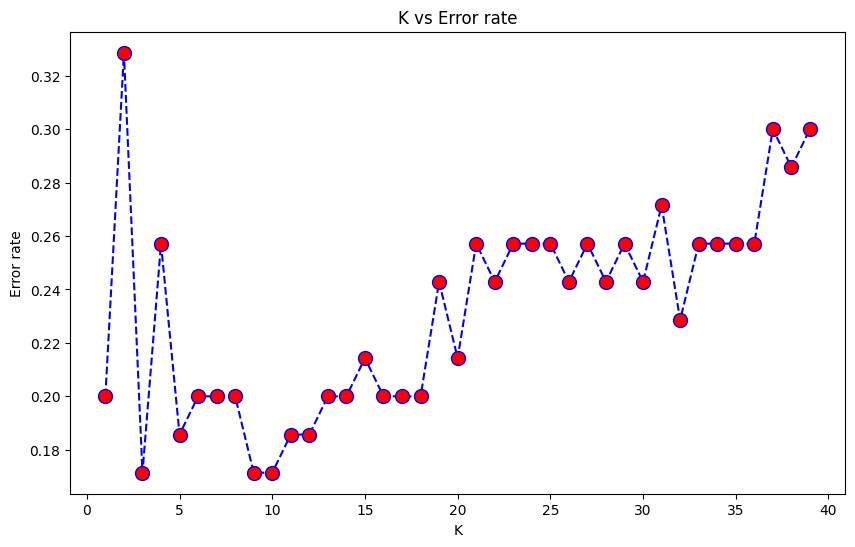
Size of the combined\_dataset: ****

For the Gene expression RNA seq data, without knowledge in the field of bioinformatics, we couldn’t remove any values.

**1) K-NEAREST NEIGHBORS:**

We implemented a basic k-nearest Neighbours (KNN) method for a cancer classification problem. After loading the dataset, unwanted columns are removed. Next, the features are scaled using the Min-Max method. A KNN classifier is trained with k=1 after the data is divided into training and testing sets. In order to evaluate the model's performance, the test set is used, and the confusion matrix and classification report are printed.

The error rate for various values of k is computed in the second section of the code, and a graphic plot is created to show how k and the error rate are related. A section illustrating the confusion matrix and classification report for a particular value of k (k=10) is also included in the code.

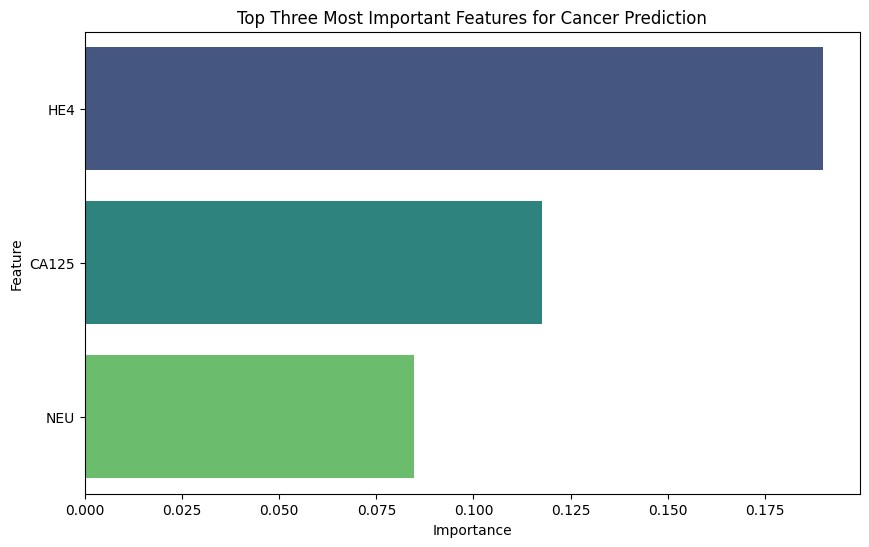
****

**2) RANDOM- FOREST:**

In this algorithm we used the scikit-learn Python package to create a Random Forest Classifier for cancer prediction. After loading the dataset, superfluous columns are removed. The data is divided into training and testing sets, and the features (X) and target variable (y) are defined. After initializing a Random Forest Classifier, a grid search is performed across combinations of hyperparameters (number of estimators, maximum features, maximum depth, and criterion) in order to identify the best set of hyperparameters using 5-fold cross-validation. After being chosen, the top model is trained on the training set and assessed on the testing set. Performance measures are computed and printed, including accuracy, precision, recall, F1 score, AUC, and log loss. Confusion matrices are additionally shown for the training and testing sets.

In addition, the code takes the trained Random Forest model's feature significance values and outputs them. A bar plot is created to represent the relative importance of the top three most significant features, which are chosen based on their importance scores. This step gives information about the features that most enhance the model's ability to forecast the future.

Overall, the code shows how to design a Random Forest Classifier, how to use grid search to optimize hyperparameters, and how to evaluate feature importance in cancer prediction.

****

**3) DECISION-TREE:**

Datasets are first combined using the method described above. The target variable is separated from the independent variables, then the data is then divided into training and testing sets. A Decision Tree model which uses the default criterion of ‘gini’ is constructed and fitted with the test data. The model’s accuracy is assessed on the test set. A Classification report and Confusion matrix is then shown to display the results. The code presents a comprehensive pipeline for training and evaluating a Decision Tree Classifier on medical data for predictive analysis.

**4) GRADIENT BOOSTING CLASSIFIER:**

Using the scikit-learn Python package, we constructed a Gradient Boosting Classifier (GBC) for cancer prediction. After the dataset is imported, the target variable is subtracted from the features. StandardScaler is used to do feature scaling after the data is split into training and testing sets. Scaled training data is used to train a baseline GBC model with predefined hyperparameters, and scaled testing data is used to evaluate the model. For both the training and testing sets, significant metrics including accuracy, log loss, area under the ROC curve (AUC), and confusion matrices are generated and supplied. The top three significant features are also shown after the feature importance’s have been determined and sorted.

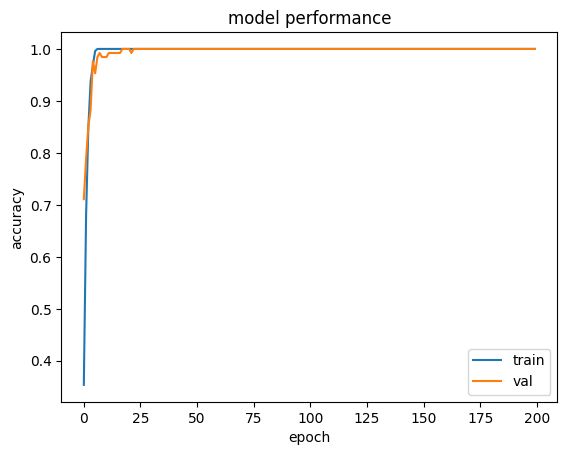
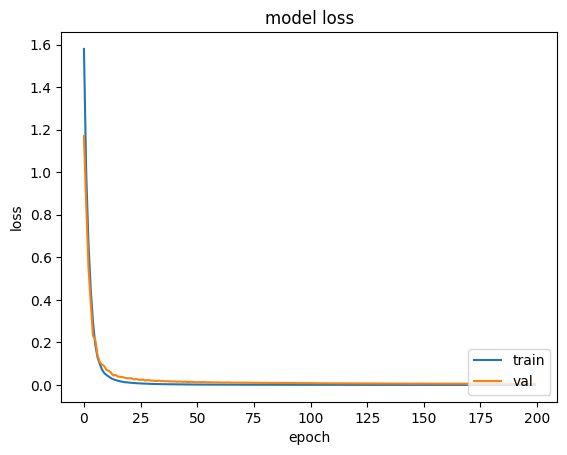
We have implemented this algo twice such that once we use the integrated dataset.

The second purpose was to implement a small prediction system using only the first OC\_Marker.csv dataset. After asking the user to enter values for each feature, the model predicts the cancer type, demonstrating the model's real-world applicability.

We also compare RF, KNN AND GBC for the combined dataset.

**5) ARTIFICIAL NEURAL NETWORKS(ANN):**

Using gene expression data as a basis, we have implemented a neural network for cancer type classification using TensorFlow and Keras libraries. A sequential neural network model is built after the label encoding, feature-label separation, and feature normalization using Min-Max scaling preprocessing steps. 40 neurons make up the input layer of the model architecture, 20 neurons make up the ReLU-activated hidden layer, and softmax activation is used in the output layer for multiclass classification. Model compilation is done using sparse categorical crossentropy loss and Adam optimizer. Training is carried out on the training set, validation is done on a different subset, and training history is recorded. Predictions are then produced after the model has been assessed using the test set. Visualizations of model accuracy and completion are plotted for both training and validation sets.



The loss graph illustrates the decrease in errors during training, while the accuracy graph tracks the model's performance on training and validation sets. Keeping an eye on these curves over epochs aids in assessing the learning process of the model and locating possible problems such as under- or overfitting.

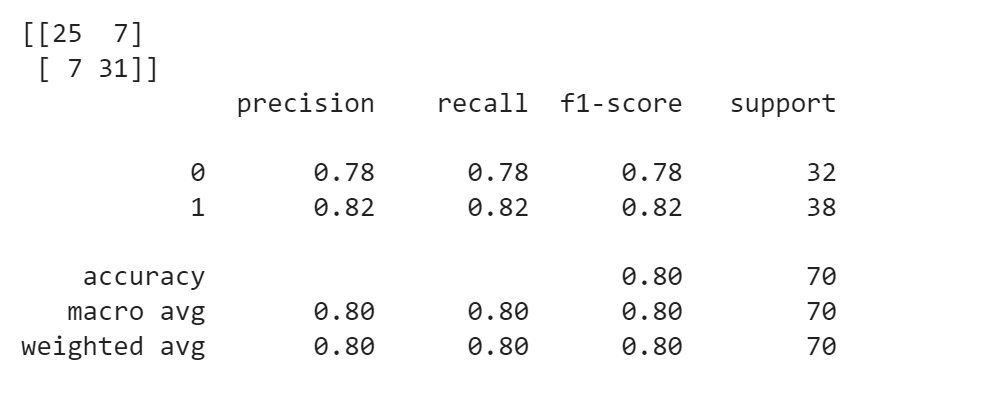
Since this model is made using gene expression data, comparing it to the other models we have implemented will not be beneficial but it still gives an amazing accuracy and shows predictability.

**RESULTS AND DISCUSSIONS**

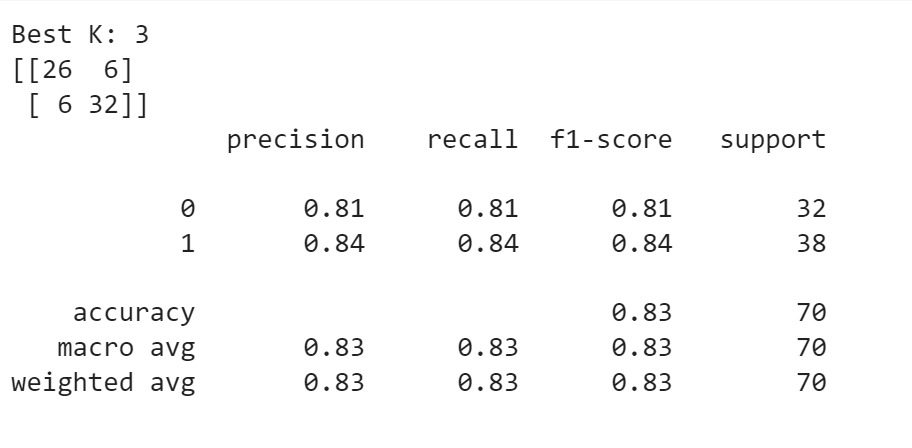
The Best Accuracies are obtained as follows:

* K-Nearest Neighbours:

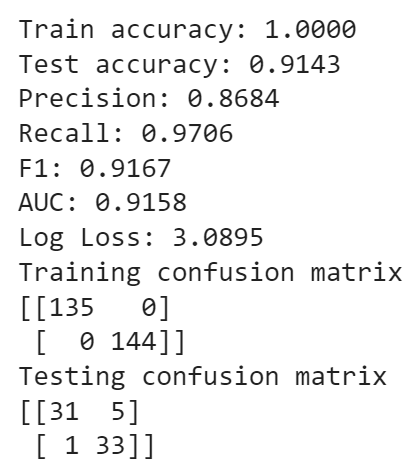
Before fitting best value of k

****

After choosing best value of k

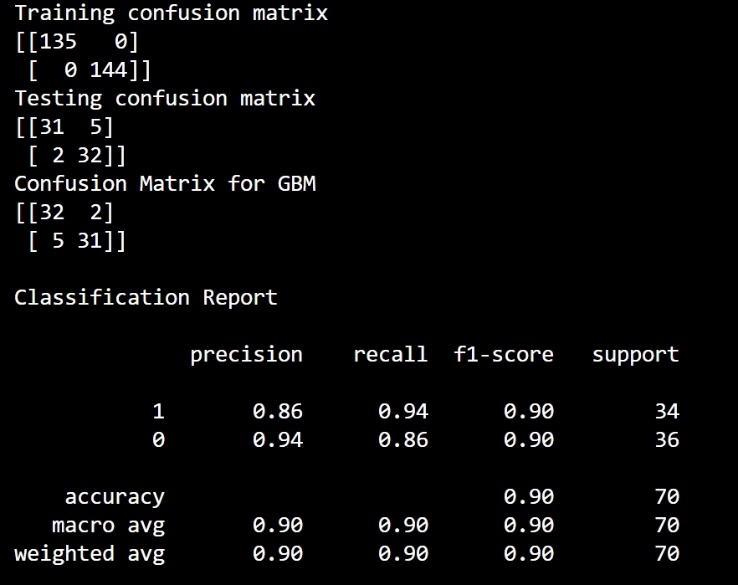


* Random Forest:

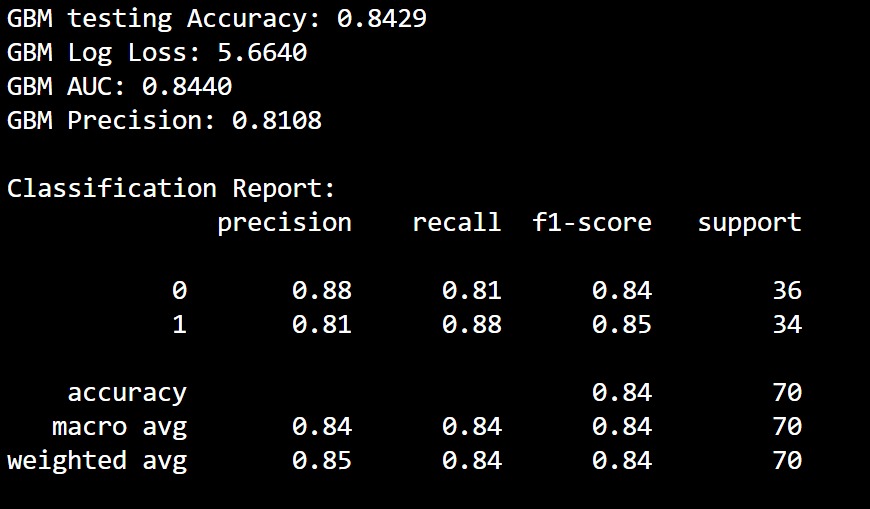
****

* Gradient Boosting Classifier:

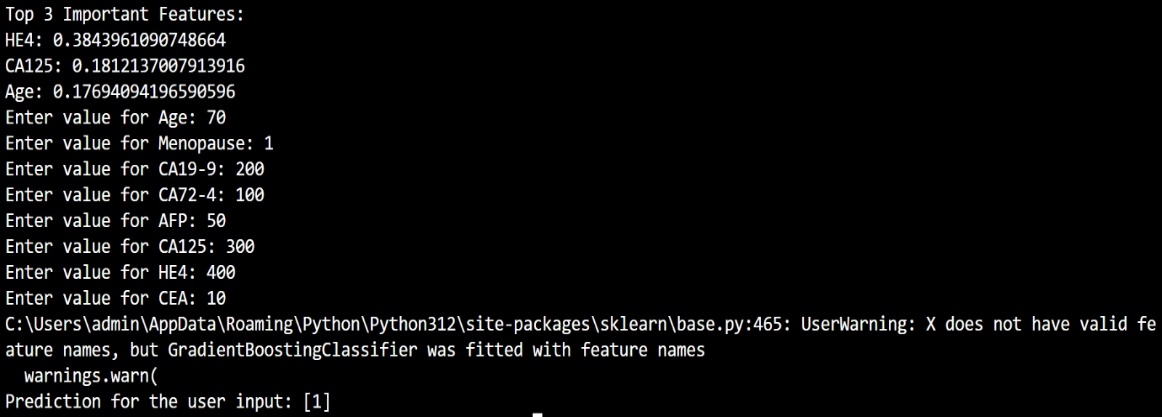
This is for the combined dataset

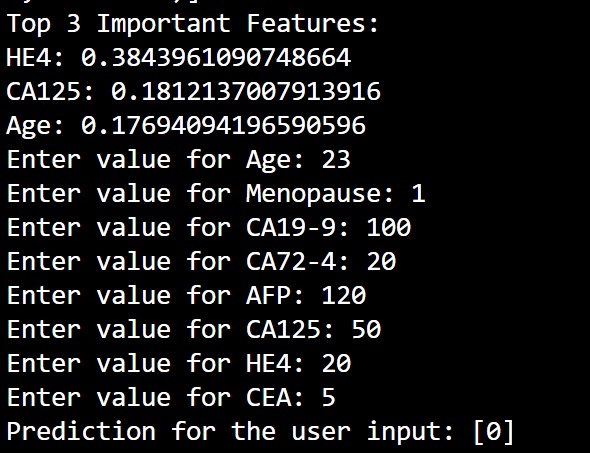
****

This is for only OC\_Marker.csv

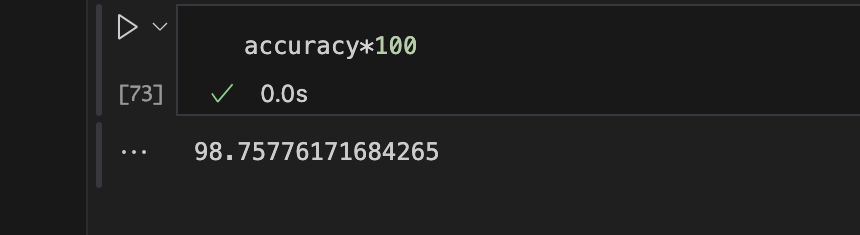


For user inputs:

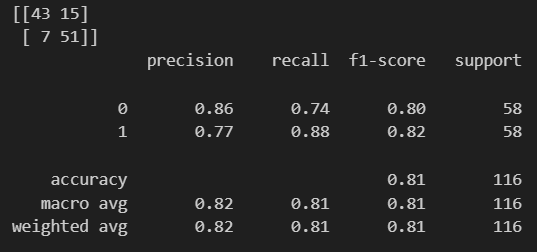


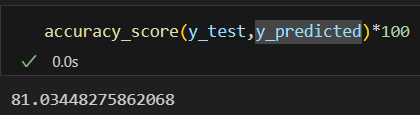


* Artificial Neural Networks (ANN):

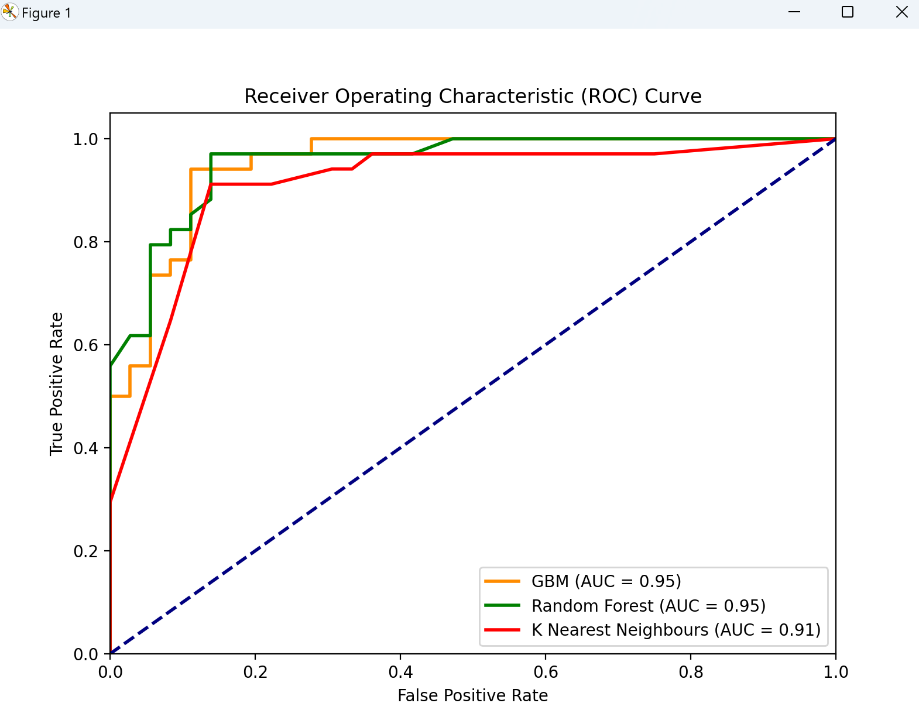


* Decision Tree:





This graph shows the comparison between GBC, RF and KNN in terms of their AUC or Area under curve. It also shows their Receiver Operating Characteristic curve.

****

Comparing the accuracies of KNN, RF, DT, GBC we can see that GBC has the greatest with 90%, RF with 86.84%, KNN with the best value of k has 83% and Decision Tree with gini criterion has 81% accuracy.

Since the dataset for ANN is widely different in nature from the others, there is no knowledge to be gained from comparing its output to the others.

On its own, ANN gives a splendid accuracy of 98% in predicting the type of cancer with the given gene expression RNA seq data.

**CONCLUSION**

In conclusion, this study employs advanced data mining and machine learning techniques, including KNN, RF amongst other models, to uncover the relationship between biological/clinical data and cancer.

The basis for further research is laid by this study, which promotes the use of larger datasets, sophisticated methodologies, and feature engineering to improve the precision and interpretability of cancer prediction models. Oncology treatment outcomes and early diagnosis could be greatly impacted by the ongoing development and application of machine learning techniques in diagnostics.

Some Applications are:

* Early Isolation of Diseases: The early disease detection capabilities of healthcare systems may be enhanced by integrating the machine learning models used in this study. Through the use of biological data, these models could help detect patterns suggestive of diseases in the early stages, allowing for early intervention and better treatment results.
* Customized Medical Care: The results of the initiative may open doors for the creation of tailored medicine strategies. Future applications could involve increasing therapeutic efficacy by customizing treatment programs based on each patient's unique biological profile, which would need an understanding of the links between biological markers and disease.
* Finding Biological Markers: The identification of biological traits that are important for illness prediction may result in the creation of new biomarkers. These biomarkers could be the focus of future investigation and the creation of diagnostic instruments, advancing our knowledge of the mechanisms underlying disease.
* Systems to Support Decisions: The accuracy of diagnoses can be improved by incorporating machine learning models into healthcare practitioners' decision support systems. These technologies may offer insightful analysis and suggestions to help doctors make well-informed decisions on patient care.
* Health Promotion Planning: Planning for public health may make use of the project's learnings. Comprehending the connections between biological information and illnesses can help to improve public health initiatives by providing guidance on health policy, resource distribution, and preventive actions.

This project also aligns with some of the United Nations Sustainable Development Goals

* SDG 3: Good Health and Well-being: SDG 3 is directly impacted by the project's prospective applications in early disease diagnosis and tailored therapy, which seek to guarantee healthy lives and promote well-being for all.
* SDG 9: Industry, Innovation, and Infrastructure: A novel strategy for tackling health issues is the application of machine learning technologies in the healthcare industry. This is in line with SDG 9's emphasis on encouraging innovation, developing resilient infrastructure, and fostering inclusive and sustainable industry.
* SDG 17: Partnerships for the Goals: Translating the project's results into useful applications requires cooperation between researchers, medical practitioners, and legislators. In order to accomplish the Sustainable Development Goals, SDG 17 highlights the significance of partnerships.

**REFERENCES**

**Dataset Links:**

[Gene Expression RNA seq](https://github.com/vappiah/Machine-Learning-Tutorials/blob/main/datasets/cancer_gene_expression.zip)

[Ovarian Cancer Clinical Data](https://github.com/martuzaiu/Ovarian_Cancer_Project)

**Literature Survey:**

[1] Debaditya Chakraborty, Elizabeth Gutierrez-Chakraborty, Cristian Rodriguez- Aguayo, Hakan Başağaoğlu, Gabriel Lopez-Berestein, Paola Amero “Discovering genetic biomarkers for targeted cancer therapeutics with eXplainable AI”

[2] LANCET DIGITAL HEALTH, Kristofer Linton-Reid, Georg Wengert, Haonan Lu, Christina Fotopoulou, Philippa Lee, Federica Petta, Luca Russo, Giacomo Avensani, Murbarik Arshard, Philipp Harter, Mitch Chen, Marc Boubnovski, Sumeet Hindocha, Ben Hunter, Sonia Prader, Joram M. Posma, Andrea Rockall, Eric O. Aboagye “End-to-End Integrative Segmentation and Radiomics Prognostic Models Improve Risk Stratification of High-Grade Serous Ovarian Cancer: A Retrospective Multi-Cohort Study”

[3] Haoxin Zhang , Meng Chi , Dongqing Su , Yuqiang Xiong , Haodong Wei , Yao Yu , Yongchun Zuo, Lei Yang “A random forest-based metabolic risk model to assess the prognosis and metabolism-related drug targets in ovarian cancer”

[4] Mingyang Lu, Zhenjiang Fan, Bin Xu, Lujun Chen, Xiao Zheng, Jundong Li , Taieb Znati , Qi Mi , Jingting Jiang “Using machine learning to predict ovarian cancer

[5] Yu, Zhong MA; Ouyang, Ling MD “Identification of prognosis-related hub genes of ovarian cancer through bioinformatics analyses and experimental verification”

[6] Kaur I, Doja MN, Ahmad T, Ahmad M, Hussain A, Nadeem A, Abd El-Latif AA. An Integrated Approach for Cancer Survival Prediction Using Data Mining Techniques. Comput Intell Neurosci. 2021 Dec 28;2021:6342226.

[7] Amirkhah et al., 2015

R. Amirkhah, U. Schmitz, M. Linnebacher, O. Wolkenhauer, A. Farazmand

MicroRNA-mRNA interactions in colorectal cancer and their role in tumor progression

[8] Weitong Huang , Hanna Suominen , Tommy Liu , Gregory Rice , Carlos Salomon , Amanda S. Barnard

[9] Lili Fan1,2†  Han Lei1†  Ying Lin1  Zhengwei Zhou1  Guang Shu3  Zhipeng Yan4  Haotian Chen3 Tianxiang  Zhang5  Gang Yin1\*

[10] Lingyan Yuan , Zhitong Bing , Peijing Yan , Rui Li , Caiyun Wang , Xueqiang Sun , Jiao Yang h, Xiue Shi , Yanying Zhang h, Kehu Yang

**Other:**

[https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9394434/#](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9394434/%23)