CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

Vishnugupthaa Ramidi | Computer Science Department | 05/03/2024

AI Campus

CSUDH

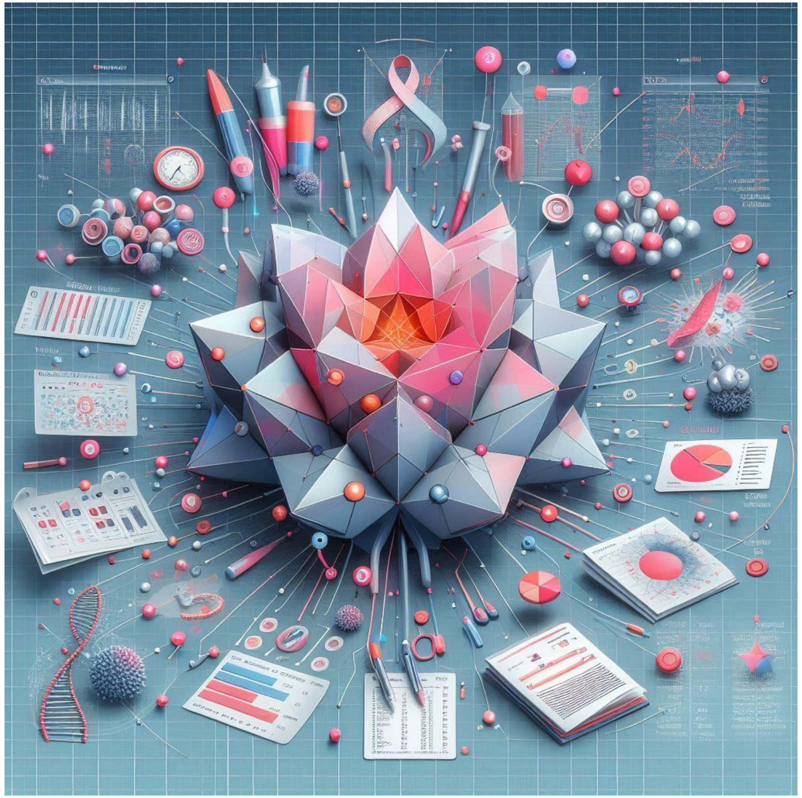
CALIFORNIA STATE UNIVERSITY, DOMINGUEZ HILLS

ABSTRACT

This research performs a multi-class classification of RNA-seq data to identify cancer subtypes, utilizing the TCGA dataset. The study applies a combination of unsupervised and supervised learning to process and analyze gene expression, revealing distinct molecular signatures. The findings enhance the understanding of cancer biology, offering potential pathways for personalized therapy and prognostic tools.

MOTIVATION AND SIGNIFICANCE

- **Importance:** This project addresses the critical need to classify cancer subtypes accurately, paving the way for targeted treatments and improved patient outcomes.
- **Impact on Cancer Genomics:** By leveraging advanced data analysis techniques, we aim to uncover molecular signatures that play a pivotal role in cancer classification, thus contributing to the broader understanding of cancer genomics.



RELATED WORK

- **Previous Studies**: A review of existing research in cancer genomics provides valuable insights and informs our approach to classification.
- **Relevance:** Understanding the findings of prior studies helps us build upon existing knowledge and develop more effective classification methods.

DATASETS

Data Collection

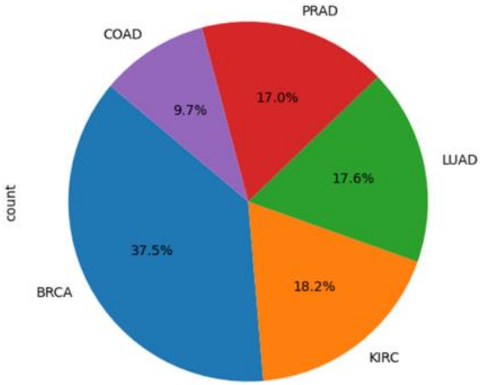
- **TCGA Pan Cancer Analysis Project:**
  - Description: A groundbreaking initiative compiling comprehensive gene expression data from diverse tumor types.
  - Significance: Provides a rich source of genomic information crucial for cancer subtype classification.
- **UCI Machine Learning Repository:**
  - Description: Additional datasets augmenting the breadth and depth of our analysis.
  - Importance: Enhances the training and evaluation phases, ensuring a comprehensive representation of cancer subtypes.

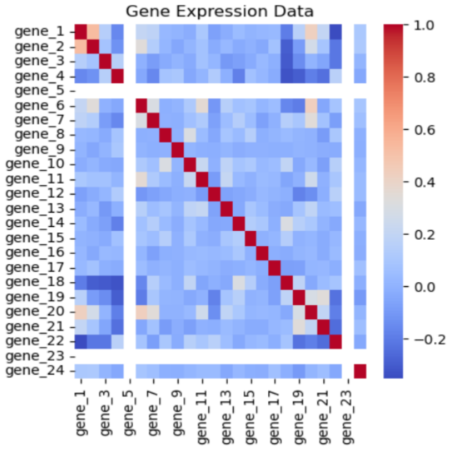
CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

DATASETS

Data Preparation

- **Steps for Data Cleaning, Preprocessing, and Merging:**
  - **Data Cleaning:** Removal of inconsistencies and incomplete entries.
  - **Preprocessing:** Normalization to standardize expression levels across samples.
  - **Merging:** Integration of datasets from different sources to create a unified dataset.
- **Importance of Data Preparation:**
  - Ensures data integrity and uniformity, facilitating accurate downstream analysis.
  - Sets the stage for dimensionality reduction, feature selection, and model development.





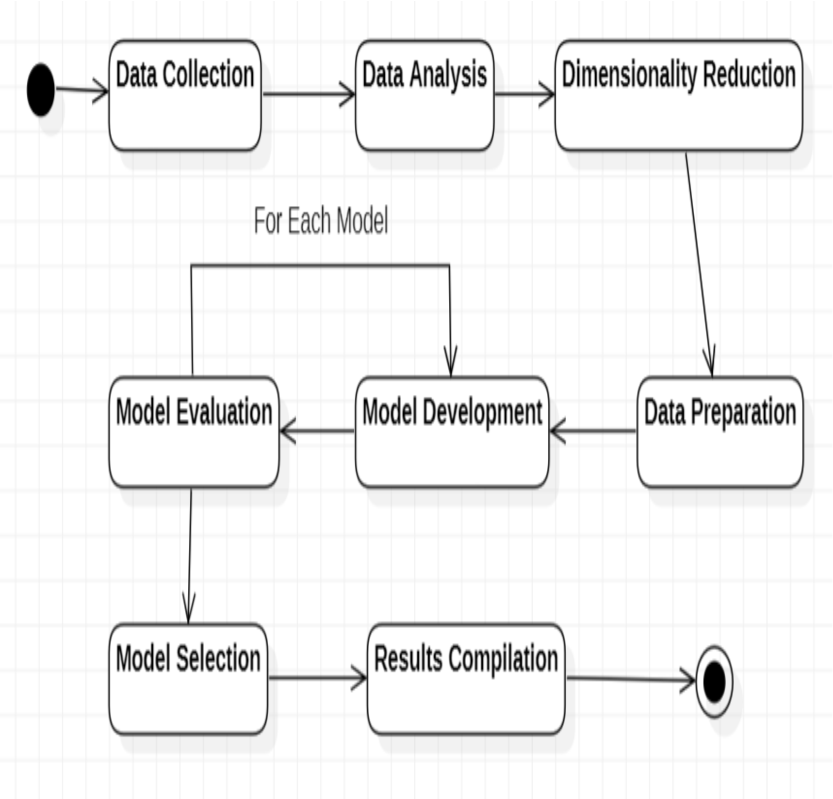
CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

SYSTEM DESIGN

System Architecture Overview:

- Sequential stages of data processing and model development.
- Guiding through data collection, analysis, dimensionality reduction, data preparation, model development, evaluation, selection, and conclusion.

1. **Data Collection:** Aggregation of gene expression data and tumor class labels.
2. **Data Analysis:** Scrutinizing dataset structure and attributes.
3. **Dimensionality Reduction:** Streamlining dataset using techniques like PCA.
4. **Data Preparation:** Refining data for model training through normalization.
5. **Model Development:** Application of classification algorithms.
6. **Model Evaluation:** Gauging model performance using metrics.
7. **Model Selection:** Choosing the most efficient model for tumor classification.

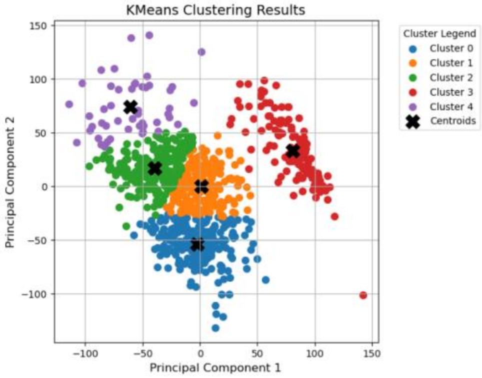


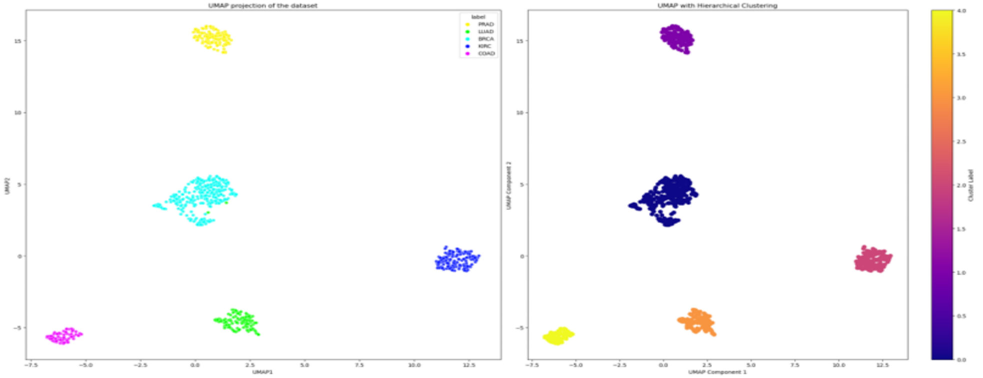
CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

UNSUPERVISED LEARNING APPROACH

Dimensionality Reduction & Clustering

- **Principal Component Analysis (PCA):**
    - Reduces dimensionality, retains informative data.
    - Includes code snippet, explained variance plot, silhouette analysis, and KMeans clustering.
- **Uniform Manifold Approximation and Projection (UMAP):**
    - Further dimensionality reduction.
    - Includes code snippet and visualized UMAP map.



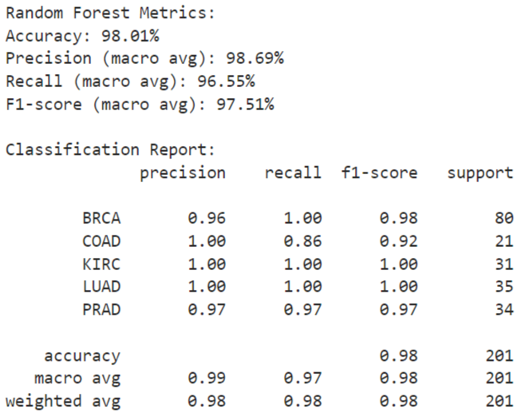


CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

SUPERVISED LEARNING APPROACH

Random Forest Classifier

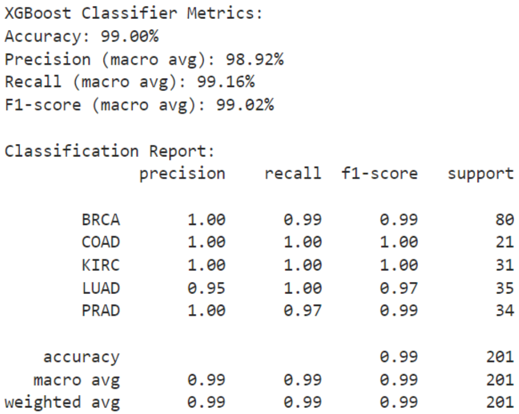
- **Ensemble Learning:** Utilizes multiple decision trees to improve predictive accuracy and control over-fitting.
- **Robustness:** Handles outliers and non-linear data effectively.
- **Feature Importance:** Provides insights into which features are most influential in predicting tumor types.



```
Random Forest Metrics:
Accuracy: 98.01%
Precision (macro avg): 98.69%
Recall (macro avg): 96.55%
F1-score (macro avg): 97.51%

Classification Report:
              precision    recall  f1-score   support

        BRCA       0.96      1.00      0.98        80
        COAD       1.00      0.86      0.92        21
        KIRC       1.00      1.00      1.00        31
        LUAD       1.00      1.00      1.00        35
        PRAD       0.97      0.97      0.97        34

    accuracy                           0.98       201
   macro avg       0.99      0.97      0.98       201
weighted avg       0.98      0.98      0.98       201
```

XGBoost Classifier

- **Speed and Performance:** Optimized gradient boosting algorithm known for its execution speed and model performance.
- **Regularization:** Includes built-in L1 and L2 regularization which helps prevent overfitting.
- **Cross-validation:** In-built routine to conduct cross-validation at each iteration.



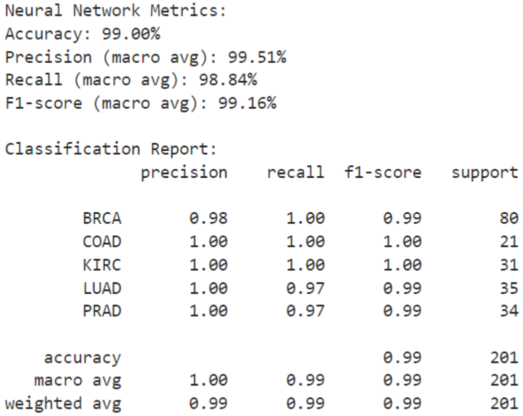
```
XGBoost Classifier Metrics:
Accuracy: 99.00%
Precision (macro avg): 98.92%
Recall (macro avg): 99.16%
F1-score (macro avg): 99.02%

Classification Report:
              precision    recall  f1-score   support

        BRCA       1.00      0.99      0.99        80
        COAD       1.00      1.00      1.00        21
        KIRC       1.00      1.00      1.00        31
        LUAD       0.95      1.00      0.97        35
        PRAD       1.00      0.97      0.99        34

    accuracy                           0.99       201
   macro avg       0.99      0.99      0.99       201
weighted avg       0.99      0.99      0.99       201
```

CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

SUPERVISED LEARNING APPROACH

Neural Network

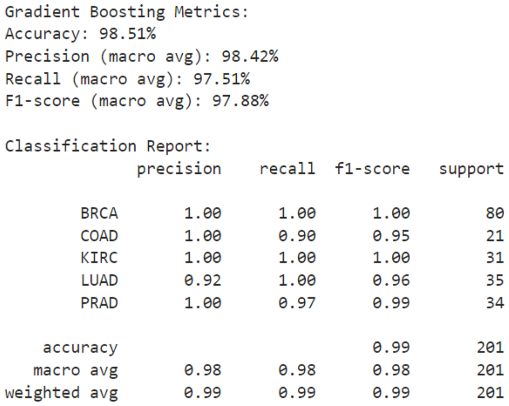
- **Flexibility:** Can model complex non-linear relationships between features.
- **Scalability:** Well-suited for large datasets and capable of learning from a vast number of features.
- **Adaptability:** Can be adjusted and fine-tuned through various architectures and hyperparameters.



```
Neural Network Metrics:
Accuracy: 99.00%
Precision (macro avg): 99.51%
Recall (macro avg): 98.84%
F1-score (macro avg): 99.16%

Classification Report:
              precision    recall  f1-score   support

        BRCA       0.98      1.00      0.99        80
        COAD       1.00      1.00      1.00        21
        KIRC       1.00      1.00      1.00        31
        LUAD       1.00      0.97      0.99        35
        PRAD       1.00      0.97      0.99        34

    accuracy                           0.99       201
   macro avg       1.00      0.99      0.99       201
weighted avg       0.99      0.99      0.99       201
```

Gradient Boosting

- **Sequential Learning:** Builds one tree at a time, where each new tree helps to correct errors made by previously trained trees.
- **Loss Optimization:** Focuses on minimizing loss function, leading to improved accuracy.
- **Handling of Missing Data:** Natively handles missing data without imputation.



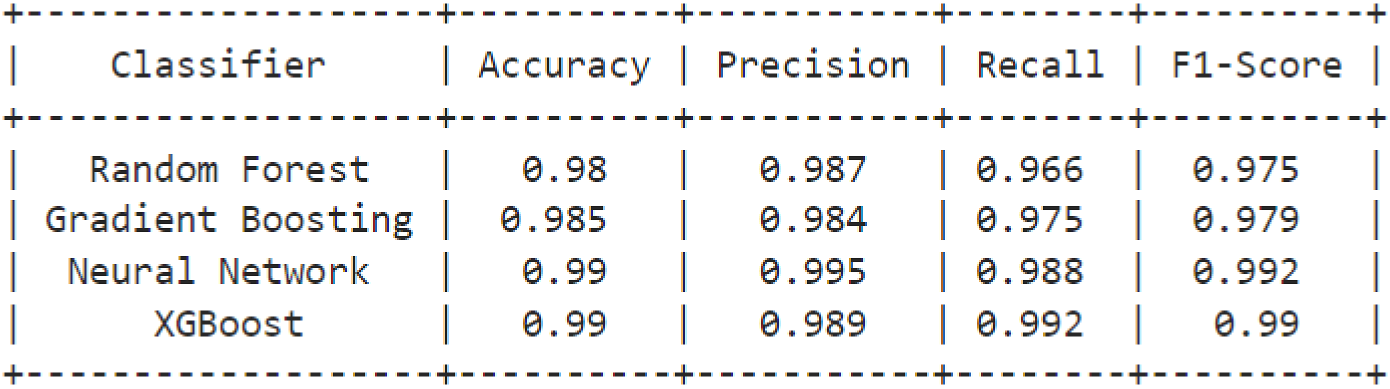
```
Gradient Boosting Metrics:
Accuracy: 98.51%
Precision (macro avg): 98.42%
Recall (macro avg): 97.51%
F1-score (macro avg): 97.88%

Classification Report:
              precision    recall  f1-score   support

        BRCA       1.00      1.00      1.00        80
        COAD       1.00      0.90      0.95        21
        KIRC       1.00      1.00      1.00        31
        LUAD       0.92      1.00      0.96        35
        PRAD       1.00      0.97      0.99        34

    accuracy                           0.99       201
   macro avg       0.98      0.98      0.98       201
weighted avg       0.99      0.99      0.99       201
```

MODEL EVALUATION AND ANALYSIS

Performance Metrics:

Key Indicators of Model Efficacy

- **Accuracy**: The model's ability to correctly classify tumor types.
- **Precision**: The ratio of true positives to all positive predictions, crucial for minimizing false positives.
- **Recall (Sensitivity):** The model's success in identifying all actual positives, highlighting its sensitivity.
- **F1-Score**: A balanced measure that considers both precision and recall, indicative of the model's overall performance.

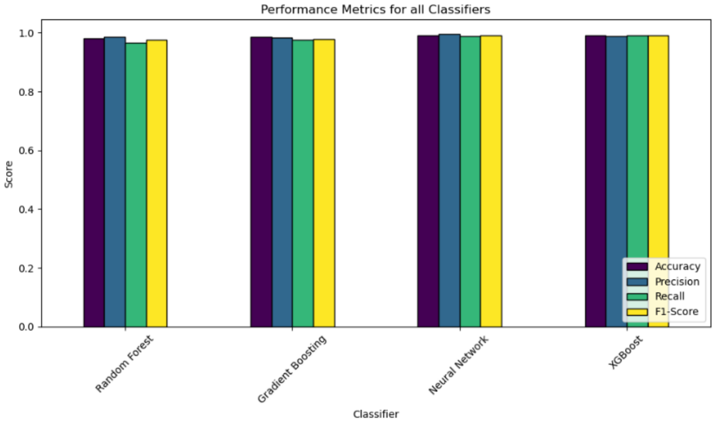


```
+-------------------+----------+-----------+--------+----------+
|     Classifier    | Accuracy | Precision | Recall | F1-Score |
+-------------------+----------+-----------+--------+----------+
|   Random Forest   |   0.98   |   0.987   | 0.966  |  0.975   |
| Gradient Boosting |  0.985   |   0.984   | 0.975  |  0.979   |
|  Neural Network   |   0.99   |   0.995   | 0.988  |  0.992   |
|      XGBoost      |   0.99   |   0.989   | 0.992  |   0.99   |
+-------------------+----------+-----------+--------+----------+
```

MODEL EVALUATION AND ANALYSIS

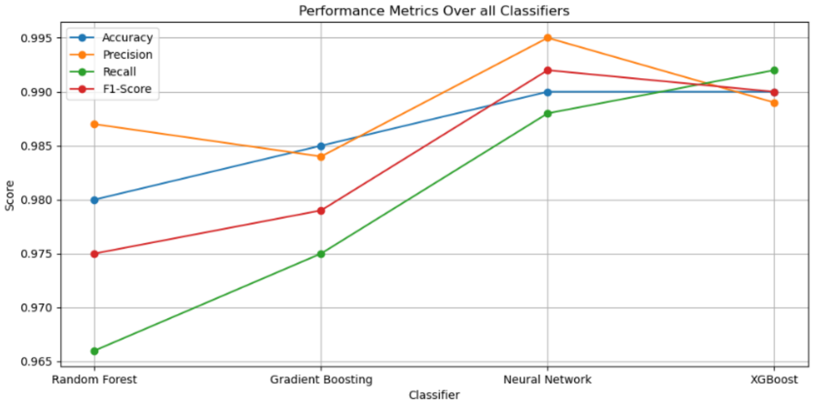
Model Performance Comparison

- Conducted a thorough comparison of RandomForest, XGBoost, Neural Network, and Gradient Boosting models.
- Evaluated using accuracy, precision, recall, and F1-score to determine each model's classification efficacy.



Insights from Comparative Analysis

- Graphical representation of performance metrics for a clear visual comparison.
- Thoughtful interpretation of metrics to understand each model's strengths and weaknesses in tumor classification.



FUTURE WORKS

- **Data Augmentation:** Enhance model robustness against overfitting.
- **Model Ensembles:** Improve performance and reliability with combined models.
- **Clinical Collaboration:** Validate models for real-world application.
- **Interpretability:** Enhance model understanding for clinical acceptance.
- **Exploring New Techniques:** Investigate novel algorithms and feature reduction methods.

Our ongoing efforts aim to refine models and contribute to cancer diagnosis and treatment advancements.

CONCLUSION

In this project, we embarked on a comprehensive journey to classify tumor types based on gene expression data. Our models, leveraging advanced machine learning techniques, exhibited high accuracy in predicting tumor classifications. The RandomForest, XGBoost, Neural Network, and Gradient Boosting algorithms each demonstrated strengths in handling complex biological data. While achieving near-perfect classification results, we critically evaluated potential overfitting and implemented strategies like cross-validation and regularization to ensure model robustness and generalizability. Our work holds significant implications for precision medicine, potentially leading to more targeted cancer treatments and improved healthcare outcomes through personalized therapy.

CANCER TYPE PREDICTION AND EXPLORATION OF TCGA DATA

THANK YOU