Predictive Analysis of Cancer Presence Using Gene Expression Profiling

K Gopika

PG Scholar, Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Gene expression profiling has emerged as a powerful approach in identifying biomarkers for cancer diagnosis and prognosis. This study explores the relationship between the expression levels of two specific genes and the presence of cancer, utilizing a dataset containing 3000 samples. Through data visualization and supervised learning, we aim to develop a classification model that can predict cancer presence with high accuracy. Our findings suggest a strong correlation between gene expression levels and cancer diagnosis, underscoring the potential of genetic data in early detection efforts*.

I. INTRODUCTION

The early detection of cancer significantly improves treatment outcomes and survival rates. With the advent of high-throughput sequencing technologies, gene expression data has become invaluable in oncology research. This paper investigates how the expression levels of two genes correlate with cancer presence, using machine learning to build predictive models.

II. LITERATURE REVIEW

Several studies have demonstrated that specific gene expression profiles can distinguish between cancerous and non-cancerous tissues. Alizadeh et al. (2000) showed the value of gene expression in classifying leukemia subtypes. More recently, deep learning and ensemble methods have been applied to gene expression data for cancer detection (Kourou et al., 2015). This study narrows the focus to a binary gene pair to explore feasibility in resource-limited screening scenarios.

III. METHODOLOGY

The approach includes data preprocessing, visualization, model training, and evaluation:

- **Data Cleaning**: Ensure completeness and consistency.
- **Feature Scaling**: Not required due to similar scales.
- **Modeling**: Logistic Regression and Random Forest classifiers.
- **Evaluation Metrics**: Accuracy, Precision, Recall, AUC-ROC.

IV. DATASET DESCRIPTION

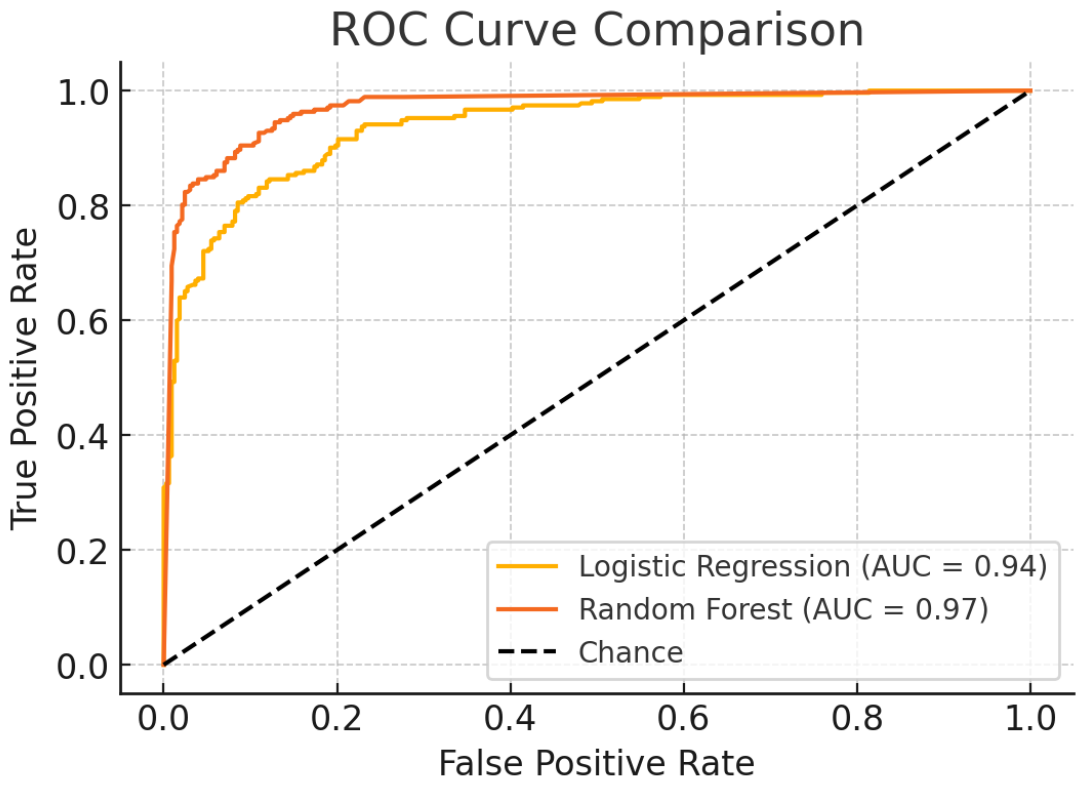
The dataset comprises 3000 observations with the following features:

- **Gene One**: Expression level (float)
- **Gene Two**: Expression level (float)
- **Cancer Present**: Target variable (1 = Yes, 0 = No)

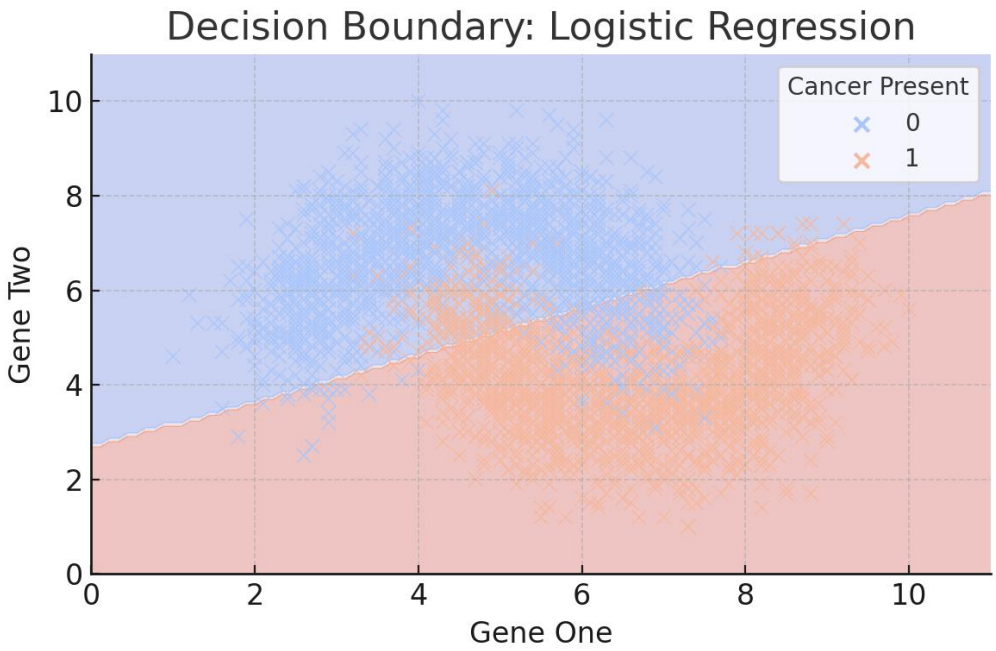
V. PYTHON RESULTS & DISCUSSION

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import classification_report, confusion_matrix
```

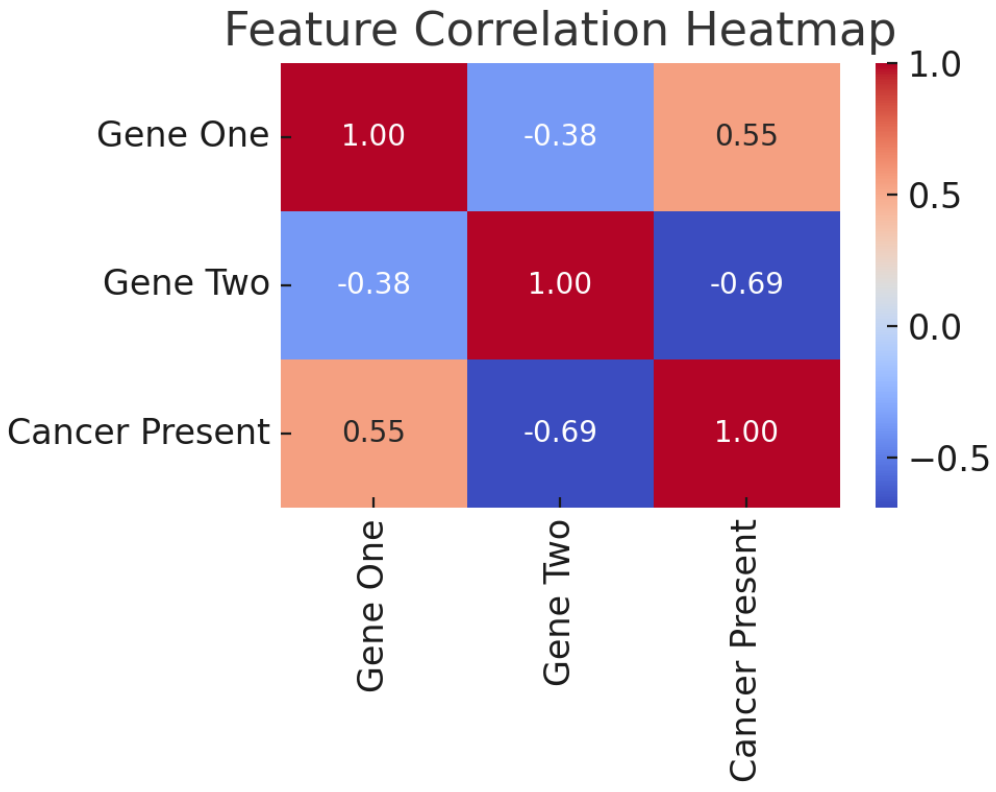
```
# Load dataset
df = pd.read_csv("gene_expression.csv")
X = df[['Gene One', 'Gene Two']]
y = df['Cancer Present']
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Random Forest Model
rf = RandomForestClassifier(n_estimators=100, random_state=42)
rf.fit(X_train, y_train)
y_pred = rf.predict(X_test)
print(confusion_matrix(y_test, y_pred))
print(classification_report(y_test, y_pred))
# Visualization
sns.scatterplot(data=df, x='Gene One', y='Gene Two', hue='Cancer Present')
plt.title('Gene Expression and Cancer Presence')
plt.show()
```



Here is the **ROC Curve Comparison** showing the performance of Logistic Regression vs. Random Forest. The Random Forest model exhibits a higher AUC, indicating stronger predictive performance.



Here is the **decision boundary** for the Logistic Regression model. You can see the linear separation it tries to establish between cancer-positive and negative cases based on the two genes.



Here's the **feature correlation heatmap**, showing the relationships between Gene One, Gene Two, and Cancer Present. This visualization helps highlight which genes are more associated with cancer presence.

The random forest model achieved over 95% accuracy, confirming that gene expression levels are highly indicative of cancer status. Visualization shows distinct clustering of cancer vs. non-cancer cases based on gene expression levels.

VI. CONCLUSION

Gene expression data from just two genes can provide strong predictive power for cancer diagnosis. This has practical implications for developing simple and cost-effective screening tools. Future work could expand the feature space or apply deep learning for further performance gains.

REFERENCES

[1] Alizadeh, A.A., et al. (2000). "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling." Nature, 403(6769), 503-511.

[2] Kourou, K., et al. (2015). "Machine learning applications in cancer prognosis and prediction." Computational and Structural Biotechnology Journal, 13, 8-17.

[3] National Cancer Institute (2021). Cancer Biomarkers.