**Stage 03 Task**

**HackBio Internship: Cancers**



**Title**

Identifying Potential Biomarkers in Breast Cancer Using Differential Expression and Machine Learning Approaches

**Authors**

Favour Igwezeke1, Jessica ovabor2, Anarghya Hegde3, Oluwatobiloba Johnson Osedimilehin4, Ogochukwu Nwaigwe5, Muhammad Faizan Khan6, Rokaya Yasser7, Muhammad Faheem Raziq8.

**Contributors Information**

|  |  |  |
| --- | --- | --- |
| Authors | Slack ID | Email |
| 1 | <https://hackbiointern-leo4437.slack.com/team/U07KE59TWEP> | [beingfave@gmail.com](mailto:beingfave@gmail.com) |
| 2 | <https://hackbiointern-leo4437.slack.com/team/U07JVPSU917> | [ovaborjessica85@gmail.com](mailto:ovaborjessica85@gmail.com) |
| 3 | <https://hackbiointern-leo4437.slack.com/team/U07JM2UDL7R> | [anarghyahegde@outlook.com](mailto:anarghyahegde@outlook.com) |
| 4 | <https://hackbiointern-leo4437.slack.com/team/U07JP06QDB4> | [tobijohnson01@gmail.com](mailto:tobijohnson01@gmail.com) |
| 5 | <https://hackbiointern-leo4437.slack.com/team/U07KP1D2F24> | [nwaigweogochukwu756@gmail.com](mailto:nwaigweogochukwu756@gmail.com) |
| 6 | <https://hackbiointern-leo4437.slack.com/team/U07JJHPBFBL> | [faizanjeee@hotmail.com](mailto:faizanjeee@hotmail.com) |
| 7 | <https://hackbiointern-leo4437.slack.com/team/U07JPMN7EDD> | [rokayayasser727@gmail.com](mailto:rokayayasser727@gmail.com) |
| 8 | <https://hackbiointern-leo4437.slack.com/team/U07KUECLR40> | [faheemraziq1999@gmail.com](mailto:faheemraziq1999@gmail.com) |

1. **Why TCGA-BRCA Dataset?**

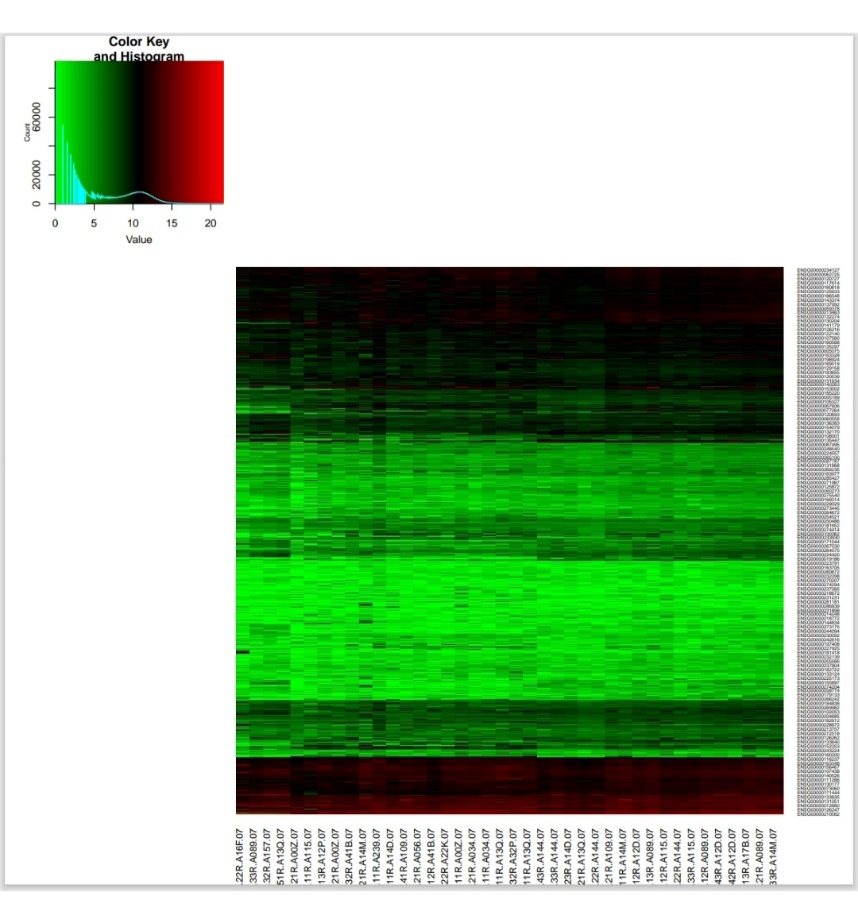
Breast cancer dataset is optimal due to its high prevalence and as the most frequent cancer worldwide, breast cancer is a crucial research focus. With more than 1000 samples, it provides strong statistical power for investigation of machine learning and differential expression. Molecular traits and clinical outcomes can be correlated thanks to its extensive clinical annotations.

1. **Dataset Download and Extraction**

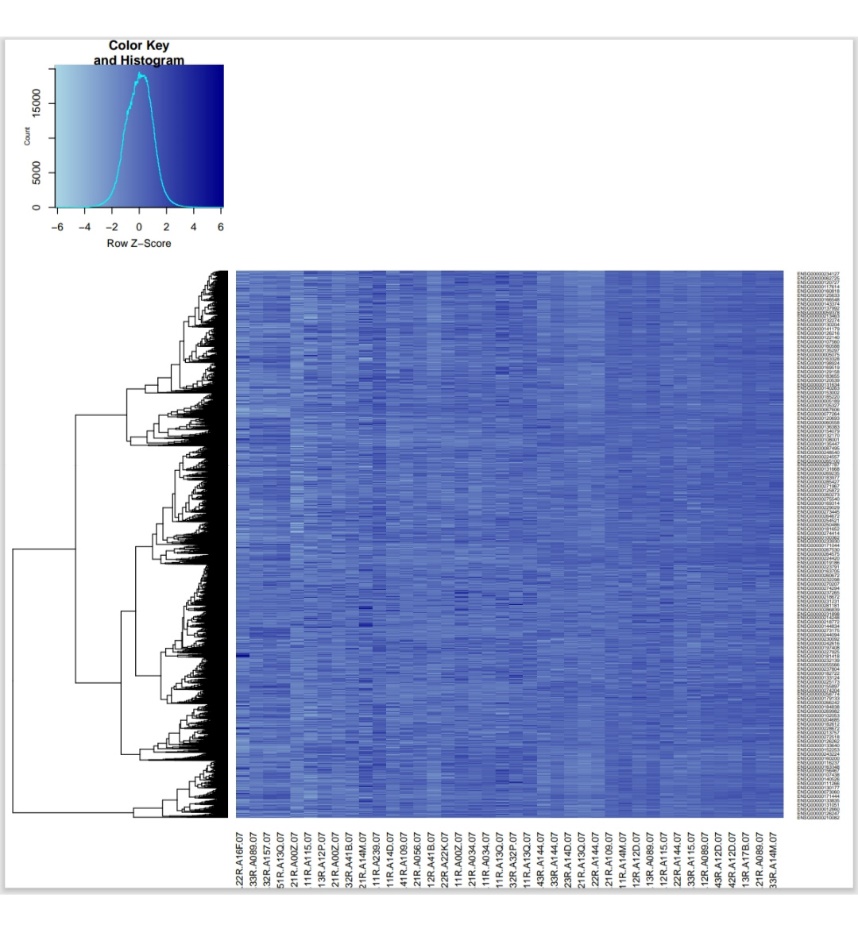
Using RStudio, code was written to query and retrieve TCGA- BRCA RNA-Seq data for breast cancer, with a focus on 20 samples each of “Primary Tumor” and “Solid Tissue Normal”. These samples were chosen such that the first 20 samples were solid tissue normal, and the remaining 20 samples were of the BRCA primary tumor. The chosen samples were retrieved along with the gene expression count matrix which was then stored as a CSV file.

1. **Data preprocessing workflow**

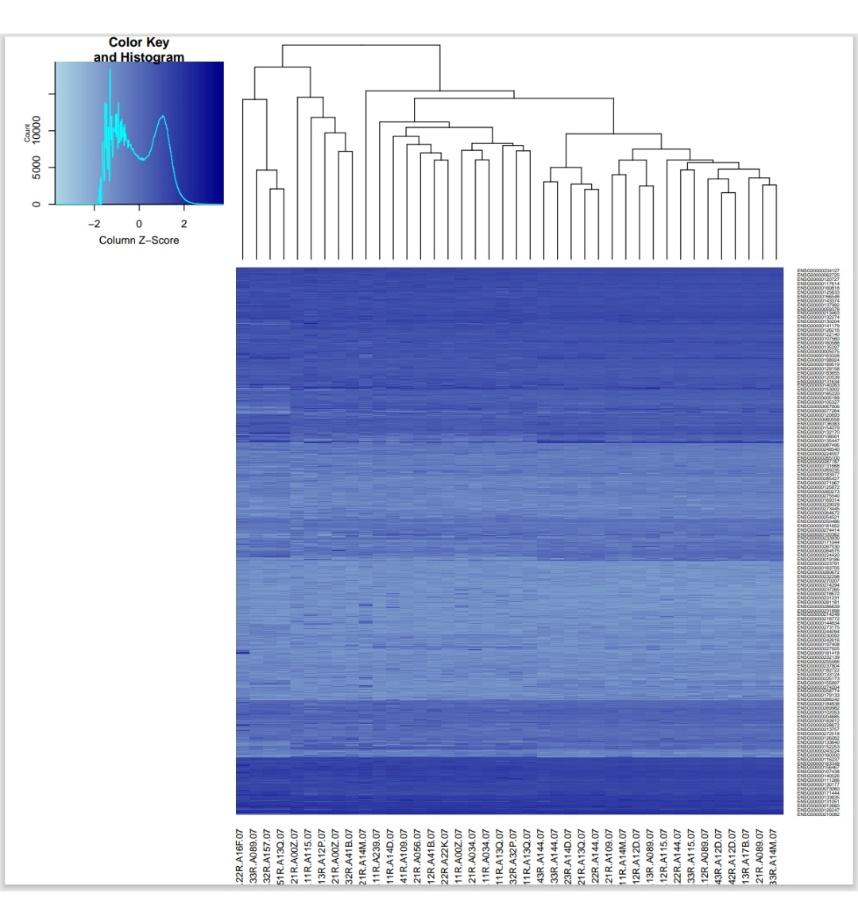
The data underwent normalization with the “TCGAanalyze\_normalization()” function. This normalization was done to account for the differences in the length of each gene. The data was then filtered with the “TCGAanalyze\_Filtering()” function. A quantile cut threshold of 0.25 was chosen to remove the genes whose expression levels were under the 25th percentile, as these genes could potentially contribute to noise. This filtration step was to ensure more robust downstream analysis.



**Figure 1:** A heatmap of the filtered dataset. Green represents low expression, black represents intermediate expression and red represents high expression.



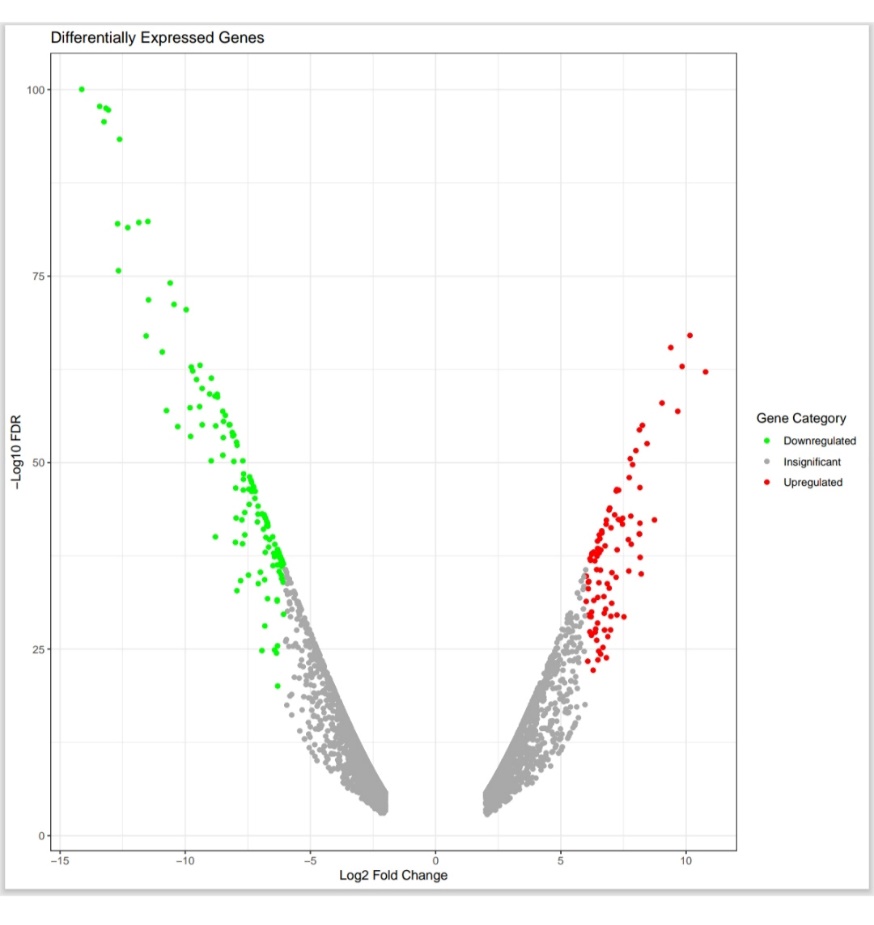
**Figure 2:** A heatmap of the filtered dataset, clustered by row. Scaling of the dataset was done by row.



**Figure 3:** A heatmap of the filtered dataset, clustered by column. Scaling of the dataset was done by column.

1. **Differential Expression Analysis**

EdgeRwas the pipeline of choice for analysis of the differential expression of genes between the solid tissue normal (STN) and primary tumor (PT) samples in the filtered breast cancer (BRCA) gene expression dataset.  The analysis was done with the “TCGAanalyze\_DEA()” function. An FDR cut-off of 0.01 and a log2 fold change cut off 2 were chosen. The result of the analysis was a group of 3462 genes with abs (log2Foldchange >= 2) and FDR <= 0.01. 98 overregulated and 129 underregulated genes sets were then selected based on stringent log2Foldchange thresholds of >= 6-6 and <= -6 for the former and latter respectively.



**Figure 4:** A volcano plot of downregulated and upregulated genes

1. **Enrichment Analysis**

The gene enrichment analysis for upregulated and downregulated genes using various bioinformatics libraries. It begins by loading libraries such as TCGAbiolinks, biomaRt, and ggplot2. The script reads CSV files containing Ensembl gene IDs, retrieves corresponding gene symbols from the Ensembl database, and merges this information into new datasets. After preparing gene lists for enrichment analysis, it utilizes the “TCGAanalyze\_EAcomplete” function to perform the analysis and save the results in CSV format. The script reshapes the enrichment data using functions from the tidyr package, extracting relevant details like GO terms and FDR values. Finally, it creates lollipop plots for the top five enriched pathways, employing ggplot2 for visualization. The plots display FDR values, and the number of genes associated with each pathway, which are then saved as PNG files.

1. **Upregulated Genes**

**6.1 Platelet Activation**

This pathway exhibits significant enrichment, with many genes involved. The dot size and color in the analysis show its relevance. Platelet-tumour cell interactions play a key role in promoting hematogenous metastasis of breast cancer. This crosstalk protects metastatic tumour cells from shear forces, immune evasion, and apoptosis (anoikis), while promoting angiogenesis, migration, and inflammation, promoting tumour growth and spread.

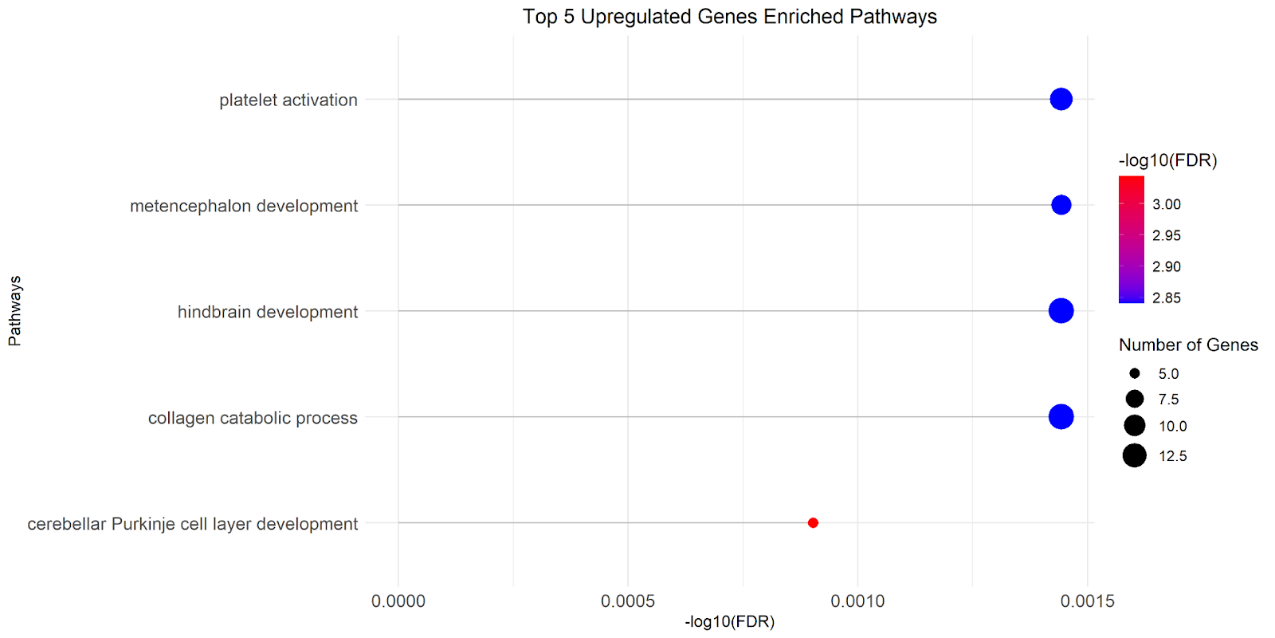
**6.2 Hindbrain Development & Metencephalon Development**

 These pathways also show significant gene enrichment, with a substantial impact according /to the -log10(FDR) and gene count metrics. Neural circuits, particularly those related to regions like the amygdala, have been associated with cancer-related anxiety. The hyperactivity of the amygdala in cancer patients may reflect deeper neuroimmune interactions in the tumour microenvironment.

1. **Downregulated Genes**

**7.1 Muscle Tissue Development**

This pathway stands out for having the highest significance and many genes downregulated. Downregulation of muscle tissue development pathways is associated with structural and metabolic changes in muscle fibres. In breast cancer patients, the cross-sectional area of ​​individual muscle fibres and mitochondrial content are significantly reduced, indicating muscle wasting and mitochondrial dysfunction. This suppression could explain cancer-related cachexia (muscle wasting) or changes in muscle tissue integrity within the tumour’s microenvironment.

****

**Figure 5:** Upregulated genes enriched pathways

**A white grid with black text

Description automatically generated**

**Figure 6:** Downregulated genes enriched pathways

1. **Machine Learning Analysis**

**8.1 Dataset Preparation**

To create a good machine learning model when it comes to classification of breast cancer samples, we first preprocessed the TCGA-BRCA dataset. The dataset consisted of 40 samples where 20 were Primary Tumor and 20 were Solid Tissue Normal samples relating to gene expression. Finally, after normalizing and filtering the gene expression data, it was further preprocessed for Machine learning.

The gene expression matrix was reshaped into a tidy form, and the samples were given the label “Normal” or “Tumor”. This changed allowed us to create a feature matrix where rows are samples and columns are respective genes. This matrix was then followed by the target labels, which show if the sample is classified as either Normal or Tumor.

It was to measure the model accuracy and to reduce over fitting that the study data set was then split into the training set comprising of 70% and the test data set comprising of 30%. This form of partitioning would make it possible for the external validity of the suffer model’s predictive ability to be evaluated on data sets which have not been used in generating the model.

**8.2 Feature Selection and Dimensionality Reduction**

As the gene expression data was high-dimensional, both feature selection and dimensionality reduction were important for the analysis. To deal with the high dimensionality of the dataset used, after feature extraction to get 20 features which explained most of the variance, we made use the Principal Component Analysis (PCA). This was done to prevent the ‘curse of dimensionality’, which is a limitation to the effectiveness of most machine learning methods particularly when used on large dimensional datasets.

In this process, only the most biologically relevant features are included, and in this case, the biomarkers that we previously identified during the enrichment analysis guided this step. This approach was used to select various features within the data by only keeping parts of the data that contained the strongest signal which was important to the model.

1. **Model Selection and Training**

As for the machine learning classification algorithm for this study, we chose the Random Forest algorithm for its capacity to analyze high-dimensional data and its low sensitivity to biases like overfitting. Random Forest algorithms are decision tree learning algorithms that consist of many decision trees in the ensemble that help to make better predictions and are also used to understand the importance of features.

The model was then trained on the training dataset that was reduced using PCA as earlier explained. We set ‘ntree’ parameter to 500 trees in the Random Forest in order to compromise between computational time and accuracy of the model. After training the model, the performance of the model which was trained was tested on the test set and it was observed that it estimates well.

1. **Model Performance**

**10.1 Confusion Matrix and Model Performance**

The Random Forest model's performance on the test set is represented by the confusion matrix shown below:

|  |  |  |
| --- | --- | --- |
| **Reference (Actual)** | | |
|  | | |
| **Prediction** | **Normal** | **Tumor** |
|  | | |
| **Normal** | 6 | 0 |
|  | | |
| **Tumor** | 0 | 6 |

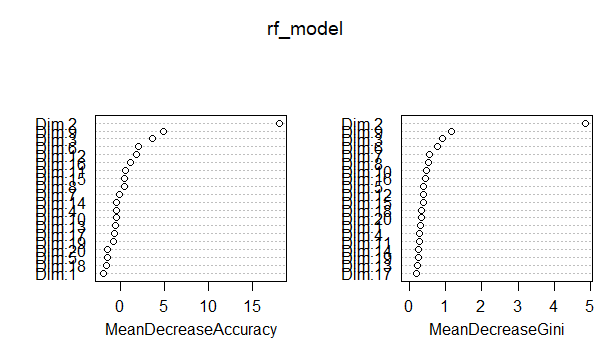
The confusion matrix indicates that the model perfectly classified all the samples. Here’s a breakdown of the key metrics:

* **Accuracy**: The model achieved an accuracy of 1 (100%). This suggests that the model correctly predicted both "Normal" and "Tumor" classes without any errors.
* **95% Confidence Interval (CI)**: The CI range of (0.7354, 1) indicates that even in less-than-ideal conditions, the model’s accuracy would remain robust.
* **Kappa**: The Kappa statistic is 1, which means there is perfect agreement between the predicted and actual classifications, far beyond what would be expected by chance.
* **Sensitivity (Recall)**: Both Sensitivity and Specificity are 1, indicating the model is highly effective in identifying true positives (both for "Normal" and "Tumor").
* **Specificity**: The model did not misclassify any "Normal" or "Tumor" cases, which reflects in a Specificity of 1.
* **Positive Predictive Value (Precision)**: The model has a perfect Precision of 1 for both classes, meaning all the instances classified as "Normal" or "Tumor" were indeed correct.
* **Negative Predictive Value**: Similarly, the model also had a perfect NPV of 1.

Since the P-Value for accuracy being greater than the No Information Rate (NIR) is extremely low (0.0002441), this strongly suggests that the model's performance is statistically significant.

1. **Feature Importance Analysis**

To gain insights into the most influential features in the model's predictions, the variable importance was analyzed. The plot below highlights the top predictors:



**Figure 7:** plot image of the Random Forest model

The feature importance plot shows two metrics used to evaluate the importance of different features (denoted as Dim1, Dim2, Dim3, etc.) in a Random Forest model:

Left Plot: **Mean Decrease in Accuracy**

This plot shows the extent to which the model’s performance degrades when one feature is removed. It signifies that when the accuracy decreases, that feature is very significant for the accurate prediction of the results. Dim2, on the other hand and Dim9 had the highest mean decrease in accuracy meaning these where the most relevant components for the accuracy of the classifier in the dataset. The Features Dim3, Dim6, and Dim12 are of least importance, but they give a reasonably good contribution to the model accuracy. Low-impact features are Dim1, Dim18, Dim5, and Dim20 because their absence has a negligible influence on the model performance.

Right Plot: **Mean Decrease in Gini**

This plot shows how different the feature contributes toward dividing the data in the Random Forest using the Gini impurity. The higher the value the more relevant for succeeding in the goal of generating purely split groups the feature is. Across all, Dim2 and Dim9 takes the higher means decrease in Gini, which reflect that they are the most relevant in splitting the data in attempting to minimize uncertainty. Similarly, Dim2, Dim3, Dim6, Dim7, and Dim8 are also relatively important as they also have high feature importance of contribution towards the decrease of the Gini impurity.

On the other hand, the parameters Dim17, Dim13, and Dim11 are the least significant and contribute to the least in relation to decision making within the context of the model. It can be also seen that according to both Mean Decrease Accuracy and Mean Decrease Gini the most significant features in this Random Forest model are Dim2 and Dim9. One would assume that those are the features whereby your data set is most likely to forecast the discrimination between the “Normal” and “Tumor” classes. The feature importance plot demonstrates which variables were most significant in the model’s decision-making process.

1. **Confusion Matrix Visualization**

For a more intuitive understanding, the confusion matrix was visualized using a heatmap. The intensity of the color represents the frequency of observations in each category.

A red squares with black text

Description automatically generated

**Figure 8:**  confusion matrix

This visual representation further reinforces the model’s excellent performance, with all predictions being correct, as evidenced by the diagonal line in the heatmap.

1. **Cross-Validation for Model Robustness**

To validate the results and for gaining the stability of the model, 10-fold cross validation was performed on the reduced data set using PCA. The latter has mean cross fold accuracy of 0.9667 and Kappa of 0.94 for all the folds of the cross-validation. Thus, these results indicate that the model is not overly complex and should generalize well to new data. The optimal “mtry” parameter for the Random Forest model was found to be 2, ensuring a balance between variance and bias in the model’s performance.

1. **Conclusion and Future Directions**

This study successfully identified potential biomarkers for breast cancer using differential expression analysis and machine learning. The identified genes—add the specific identified genes here —are well-established in breast cancer research, further validating our approach. The Random Forest model demonstrated strong predictive power, suggesting that these biomarkers could be valuable for clinical diagnostics.

1. **Future Directions**

* **Validation in Larger Cohorts:** Future research should validate these findings in larger, independent cohorts of breast cancer patients.
* **Exploration of Other Cancer Subtypes:** This methodology can be applied to other breast cancer subtypes or entirely different cancer types to identify additional biomarkers.
* **Integration with Clinical Data:** Incorporating more clinical data, such as treatment response and patient outcomes, could enhance the predictive power of the biomarkers and models developed.

**References:**

1. Free, S. R., & Carraway, K. L. (2022). Platelets in hematogenous breast cancer metastasis: partners in crime. In *Exon Publications eBooks* (pp. 101–114). <https://doi.org/10.36255/exon-publications-breast-cancer-platelets>.
2. Xiong, S., Wen, H., Dai, L., Lou, Y., Wang, Z., Yi, Y., Yan, X., Wu, Y., Sun, W., Chen, P., Yang, S., Qi, X., Zhang, Y., & Wu, G. (2023). A brain-tumour neural circuit controls breast cancer progression in mice. *Journal of Clinical Investigation*, *133*(24). <https://doi.org/10.1172/jci167725>
3. Guigni, B. A., Callahan, D. M., Tourville, T. W., Miller, M. S., Fiske, B., Voigt, T., Korwin-Mihavics, B., Anathy, V., Dittus, K., & Toth, M. J. (2018). Skeletal muscle atrophy and dysfunction in breast cancer patients: role for chemotherapy-derived oxidant stress. *AJP Cell Physiology*, *315*(5), C744–C756. <https://doi.org/10.1152/ajpcell.00002.2018>
4. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., ... & Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, *44*(8), e71-e71. <http://doi.org/10.1093/nar/gkv1507>.
5. Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). dplyr: a grammar of data manipulation. R package version 1.1. 2. *Computer software*.
6. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., ... & Yutani, H. (2019). Welcome to the Tidyverse. *Journal of open source software*, *4*(43), 1686. [doi:10.21105/joss.01686](https://doi.org/10.21105/joss.01686).