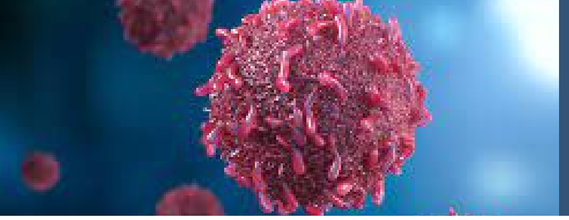
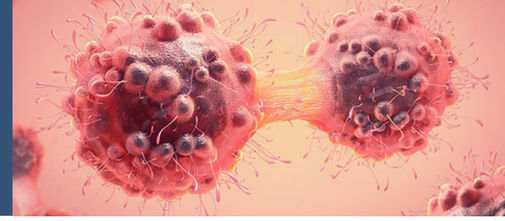


Binary Cancer Cell Classification

Malignant or Benign

Cheeto Team (32) - Mukta Jaiswal, Nathan D'Cruz, Jonathan Yeung, Nguyen Ho



01. Introduction and Addressing the Problem

According to scientific predictions, 2 million people will be diagnosed with cancer and **609,000 will die** this year. Patient survival is correlated to early detection and identification of the cancer, but some types of cancer can be hard to correctly diagnose. This is where our project aims to fill the gaps.

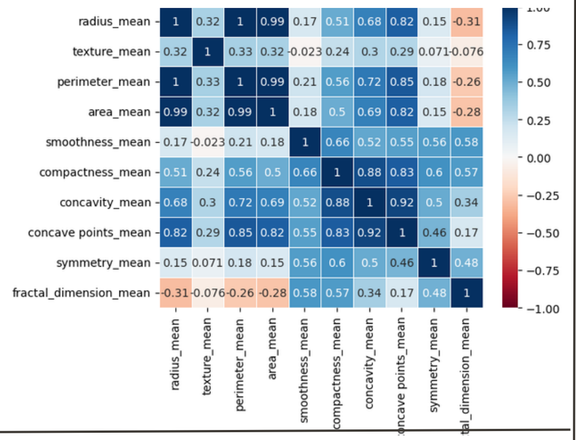
02. Dataset and Methods

Dataset Information:

1. The dataset includes **570 records of tumors**
2. Each record includes up to **30 attributes** describing the appearance of the tumor with characteristics
3. **63%** of dataset is **benign** tumors and **37%** for **malignant** tumors

Dataset Exploratory Analysis:

1. **Dropped** redundant or unnecessary columns (e.g., id column).
2. Generated **correlation matrix** to identify closely correlated columns (e.g., radius, perimeter, and area) and either dropped or combined them to avoid multicollinearity issues.
3. **Normalized the data** to bring extreme values between 0 and 1, reducing prediction model errors.
4. **Binarized the M and B values** of the prediction column (diagnosis column) as 1 and 0 for analysis.



Model Types used:

1. Logistic regression
2. Feed-forward neural network
3. Support Vector Machine

03. Experimental Results and Findings

We judged our models on **accuracy, F-Score, and mean-squared error**. We chose to consider F-Score because our dataset has slightly more benign samples than malignant. However, in most cases the F-Score was close to the accuracy metric, so we believe that the sample imbalance was not a major factor in this case.

We performed **K-fold cross-validation** to assess the generalization of our model and yielded a very high average accuracy for this model. This ensures that the model performs consistently on unseen data and **guards against overfitting or under-fitting issues**.

1. We classified the data points into categories which gave us a probability estimate of our result. — **Logistic Reg.**
2. The nodes in the feed-forward neural network were processed with a **sigmoid/logistic activation function**. — **Feed Forward NN**
3. The use of **normalization** and increased the accuracy of the SVM by roughly 1%. — **Support Vector Machine**

Model	Accuracy	MSE	F-Score
Logistic Regression	0.9578	0.0422	0.9470
Feed-Forward Neural Network	0.8875	0.1124	0.8520
Support Vector Machine (no oversampling)	0.9648	0.0352	0.9490

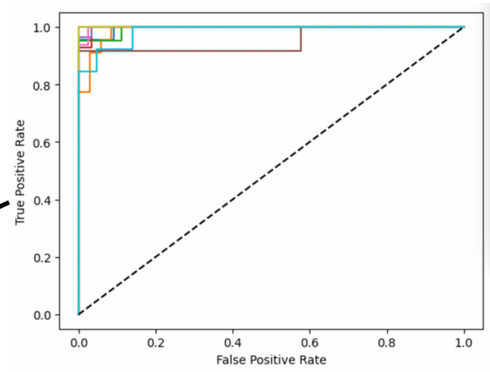


Figure 3: ROC Curve for Logistic Regression

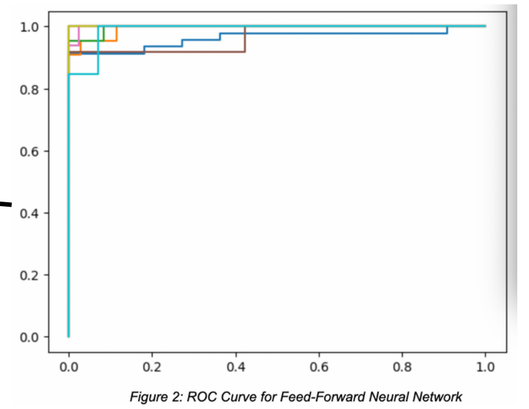


Figure 2: ROC Curve for Feed-Forward Neural Network

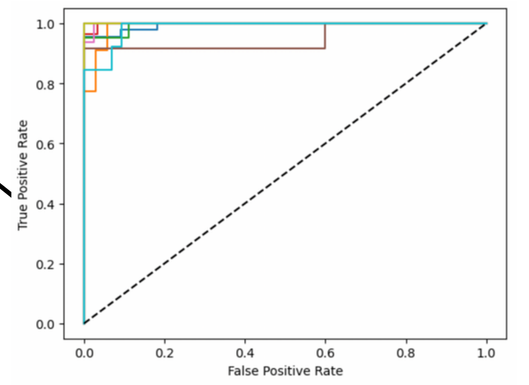


Figure 5: ROC Curve for Support Vector Machine (no oversampling)

04. Conclusion and Key Takeaways

In conclusion, we found the Support Vector Machine (SVM) to be the best choice in classifying tumors **based on the average accuracy and F-score.** Compared to the other models we tested, it also had the **lowest difference in bias and variance** between the training and testing data.

Although we found that our SVM was slightly overfitting, with a r**elatively low bias (3.28 vs 3.52) and high variance (-2.95 vs -3.17)**, we could reduce this **by performing hyperparameter tuning**. Based on our implementation of **hyperparameter tuning** for our SVM model, we found the optimal parameters to use to prevent overfitting were - a linear kernel, with a gamma value of 0.1, and a regularization parameter strength of 10.

We also chose to **not oversample our SVM** because it appeared to artificially increase the SVM's accuracy. In the future, a more reliable dataset with more samples may be required. Additionally, in the future, we can instead create a model that classifies tumors as malignant or benign **based on imagery**.

05. References

https://www.westga.edu/academics/research/vrc/assets/docs/scatterplots_and_correlation_notes.pdf

https://www.cdc.gov/cancer/dcpc/research/update-on-cancer-deaths/index.htm

https://seer.cancer.gov/statfacts/html/common.html

https://canvas.ucdavis.edu/courses/786744