**Background**

* **Brain Tumor** is an abnormal growth of tissues in the brain and is a fatal health issue. Detection of tumors in the brain would aid in starting the treatment early and prolonging a patient's life.
* **Magnetic Resonance** (MR) images are used to identify a tumor as it shows clear boundaries among the unaffected and anomaly areas.
* In this project, I have proposed a model to detect the presence of brain tumors from MR scans.
* The detection from the initial scans happens in two stages – **feature extraction and tumor detection.**
* Feature extraction and **reduction are performed using Principal Component Analysis (PCA).**
* Two Machine Learning models**, Support Vector Machine (SVM) and Random Forest (RF)** are implemented for the detection.
* This study results in an accuracy of **97% with the SVM** classifier with an training time of just **3.8** seconds and **RF classifier gave a 96% accuracy** with an training time of **15.3 seconds**.

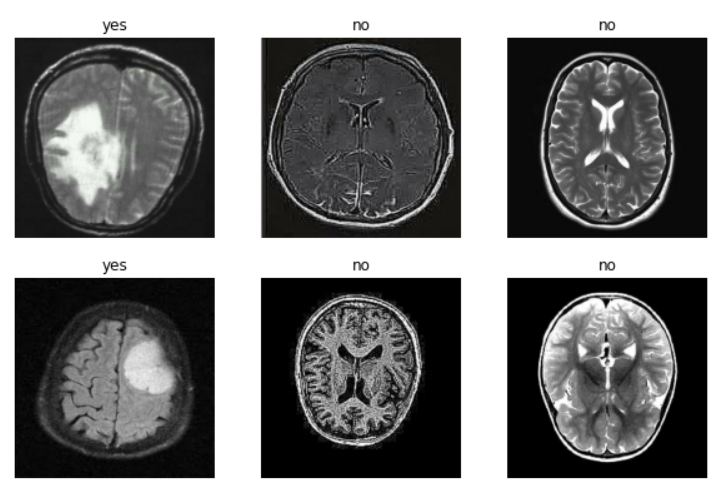


* The dataset consists of 3000 MR scans – 1500 images with tumor and 1500 non-tumor images.
* The entire data was first preprocessed in 2 steps:

Converted to greyscale and resized to 256 x 256.

* Table 1 - there is a class balance between tumor and non-tumor classes.
* The open-source data was obtained from Kaggle.

|  |  |  |
| --- | --- | --- |
|  | **Train** | **Test** |
| Tumor | 1187 | 313 |
| Non-tumor | 1213 | 287 |

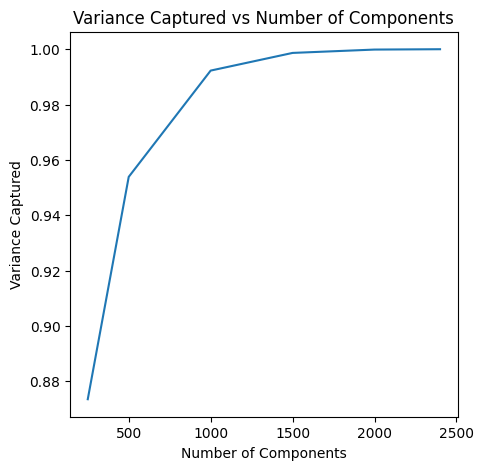
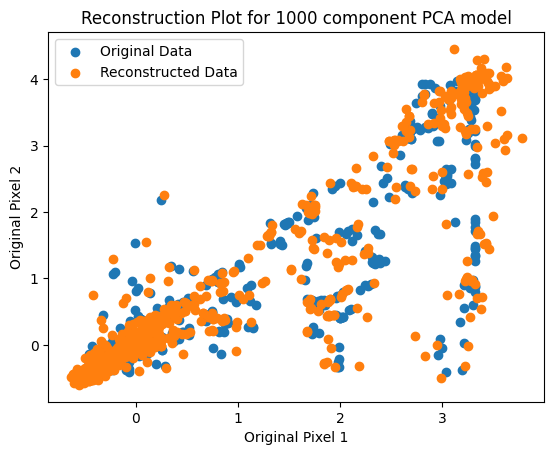


*Fig 1:* Sample MR images from dataset



The proposed method has 4 steps starting from input data to the detection of tumor.

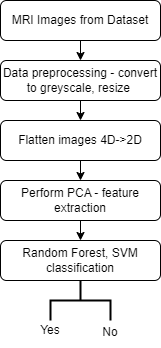
1. The initial images are converted to greyscale and resized to 256x256.
2. The 4D tensors of shape (3000, 256, 256, 1) are **flattened** to 2D tensors of shape (3000, 65536). In the resulting dataset, each row is an image and each column is a pixel starting from top left to bottom right.
3. PCA is applied to reduce the dimensionality of the training and testing data from **65536** dimensions to **1000 dimensions**.

* **1000 principal components captured 99% of the variance.**
* **Reconstruction plot** created from the transformed data shows that 1000 **principal components** precisely **capture the main patterns** and **variation** from the original data (see Fig 3).

1. Train **Random Forest (RF) classifier and SVM classifier** on the reduced dimensional data (only training set) of shape (2400,1000). The resulting models are then used to test using the testing data of shape (600, 1000) and a predictive accuracy was obtained using the original labeled outputs.

*Fig 2: Proposed Model for PCA-RF-SVM*

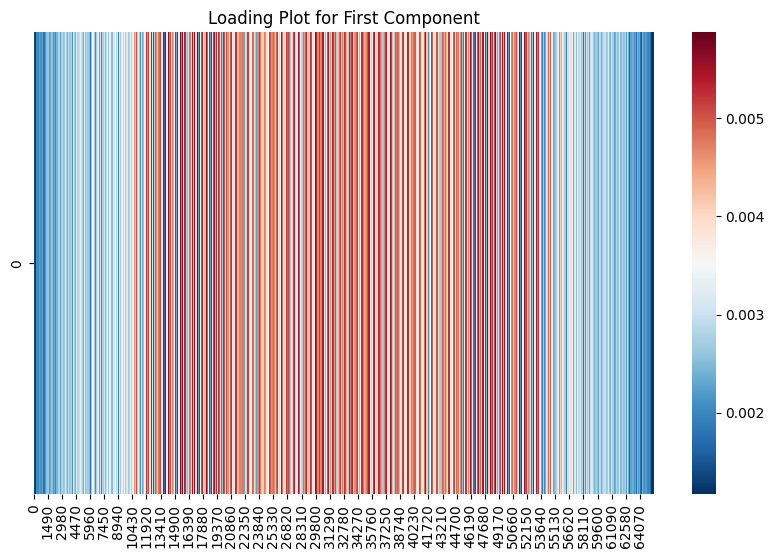




PCA intuition

The first few components of the PCA model are analyzed to understand how the model might be working with the original data. Loading plot analyses:

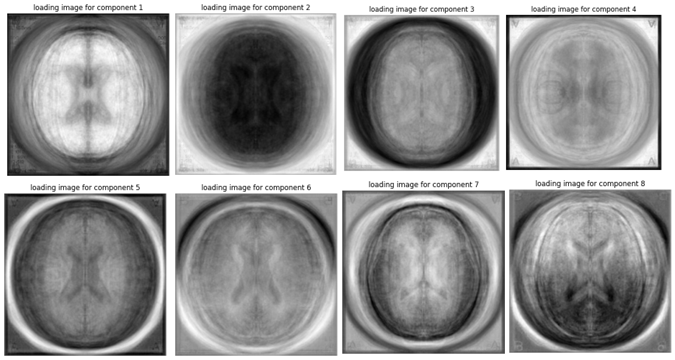
* The P1 heatmap represents how each pixel contributes to the first principal component.(Fig 4)
* **Blue indicates lower/negative loading values, while red indicates higher/positive loading values**.
* The x-axis of the heatmap is labeled with numbers ranging from 0 to **65536**, representing each flattened pixel of the tumor/non-tumor images.
* According to the p1 loading plot, **the first and last ~1000 pixels**, **center ~28000-35700 pixels** contribute significantly to the first principal component.
* This could be related to features that differentiate tumor from non-tumor regions in the images. The pixels in the center is where the brain MRI is predominantly present.



*Fig 4: P1 Loading plot as heatmap*

Another way to understand a large number of loadings(65536) is to display them as images in greyscale.

*Fig 5: P1-P8 Loading plots as images*

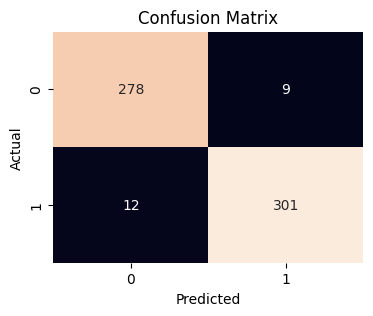


1. Darker intensities (**black**) represent – extreme **negative** **loading**
2. Lighter intensities (**Grey**) represent - lower **negative and positive** loadings
3. **White** represents – **positive** loadings
4. The **outer layers of the circle** represent the **top and bottom pixels of the MR scan** and the inner layers represent center pixels.
5. These loading plots illustrate that each component picks up different areas of pixels to contribute to that component.

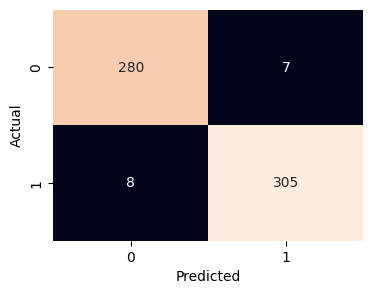
**RF and SVM**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Accuracy** | **F1** | **Execution time (Seconds)** |
| **RF Classifier** | 96% | 0.96 | 15.3 |
| **SVM Classifier** | 97% | 0.98 | **3.8** |

**Random Forest Confusion matrix**

****

**SVM Confusion matrix**

****