Malaria Detection Through RBC Images

Vivek Jain : 2021218
Manan Garg : 2021163
Tushar : 2020478

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

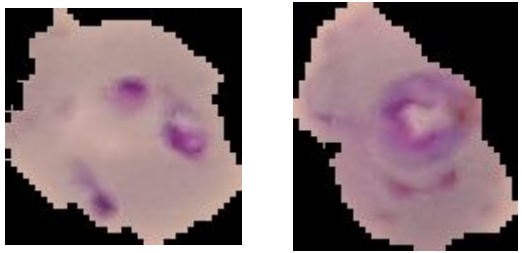
- Malaria claims over 400,000 lives each year, underlining the urgent need for an efficient and accessible diagnostic method to ensure timely treatment and reduce mortality rates
- In many resource-limited areas, access to accurate malaria diagnosis is severely restricted, hindering prompt medical intervention. Automation through machine learning can bridge this gap, providing a scalable and cost-effective solution.
- Recognizing the transformative potential of machine learning, the project aims to harness its capabilities to significantly enhance the accuracy and speed of malaria detection, revolutionizing the diagnostic process and ultimately saving lives.

Literature Review

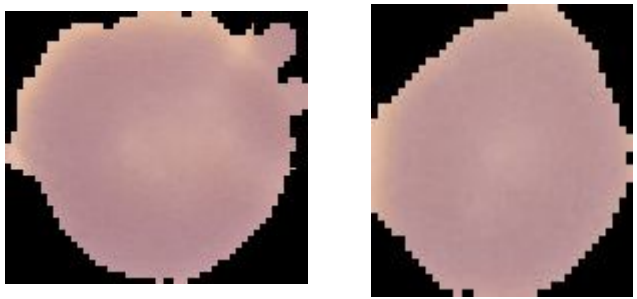
- **PERFORMANCE COMPARISON OF MACHINE LEARNING ALGORITHMS FOR MALARIA DETECTION USING MICROSCOPIC IMAGES** : - The study, led by Dr. Saiprasath G and his research team, emphasizes the use of classical machine learning algorithms, including AdaBoost, Decision Tree, KNN, and Random Forest, to analyze patches of images for malarial parasite presence.Notably, their research highlights the exceptional effectiveness of Random Forest, achieving an impressive accuracy of 96.5%, in automating malaria parasite detection within blood smear images, showcasing the potential of machine learning in advancing malaria diagnostics.
- **Melanoma Detection in Dermoscopic Images using Global and Local Feature Extraction :** Dr. JC Kavitha and her research team introduced an effective approach for identifying melanoma in dermoscopic images. Utilizing Support Vector Machine with Radial Basis Function (SVM-RBF) and K-Nearest Neighbor (KNN) for classification, their experiments demonstrated that leveraging the local texture feature descriptor SURF in conjunction with SVM-RBF yielded superior results. Specifically, they achieved a sensitivity of 86.2%, specificity of 88.4%, and an impressive overall accuracy of 87.3%, underscoring the efficacy of their proposed methodology in accurately detecting melanoma.

Dataset:

- The data is titled as NIH Malaria Dataset
- Dataset comprises of 27558 images of uninfected (13779) and infected (13779) with plasmodium in total.
- The segmented red blood cell patches have 3 channels (RGB) and size range of 110-150 pixels.
- Samples with plasmodium present in them are positive (labeled as 1) else negative (labeled as 0)
- Other than plasmodium staining impurities is also present in it.
- It is divided into two sets: training test and testing set with a ratio of 75:25.



Infected



Uninfected

Details regarding features and preprocessing

Three main categories of features were extracted from the given RBC images.

The Three categories were :

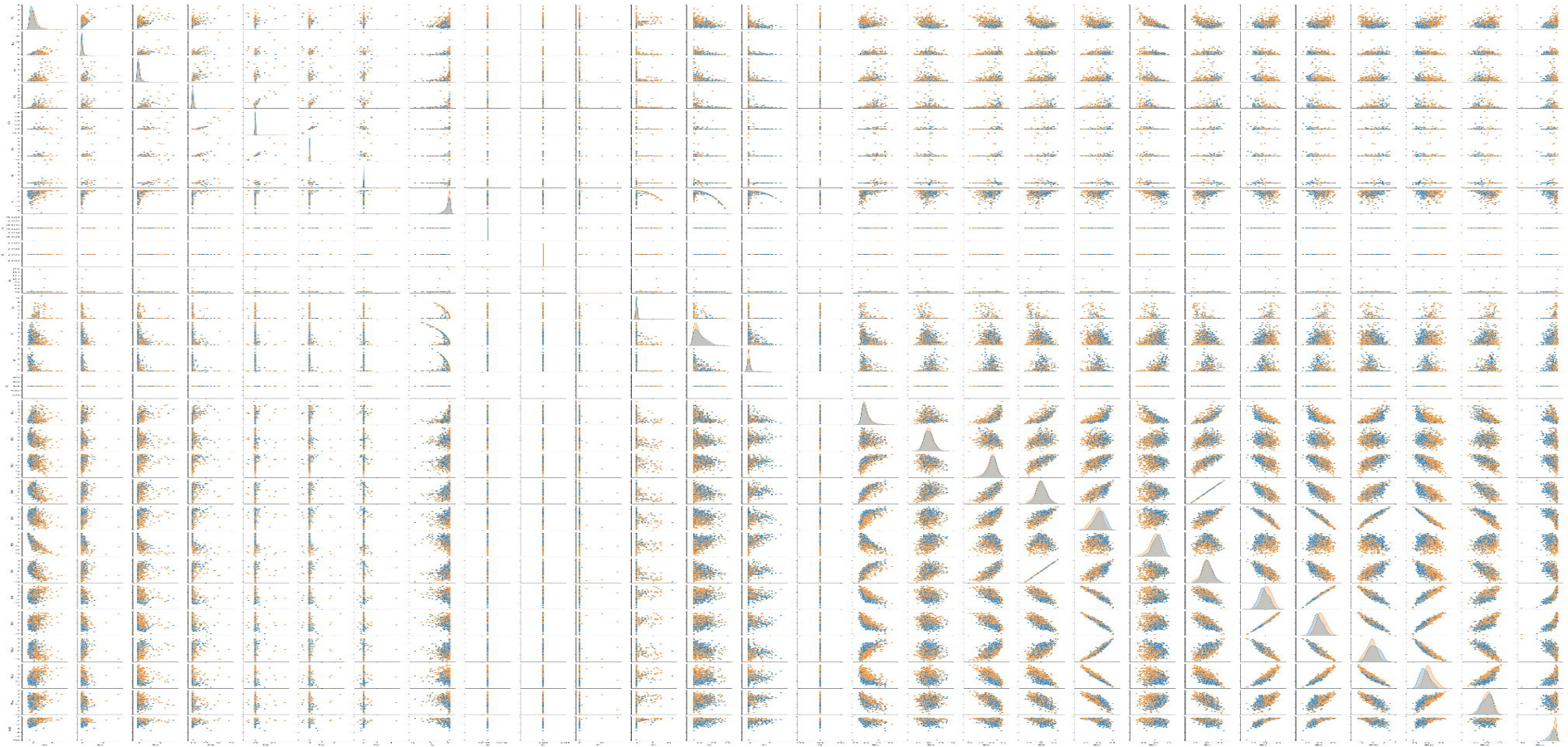
1) HUmoments for extracting shapes (7 subfeatures)
2) Haralick for extracting textures (13 subfeatures)
3) Histogram for extracting colors (8 subfeatures)

Total features = 28

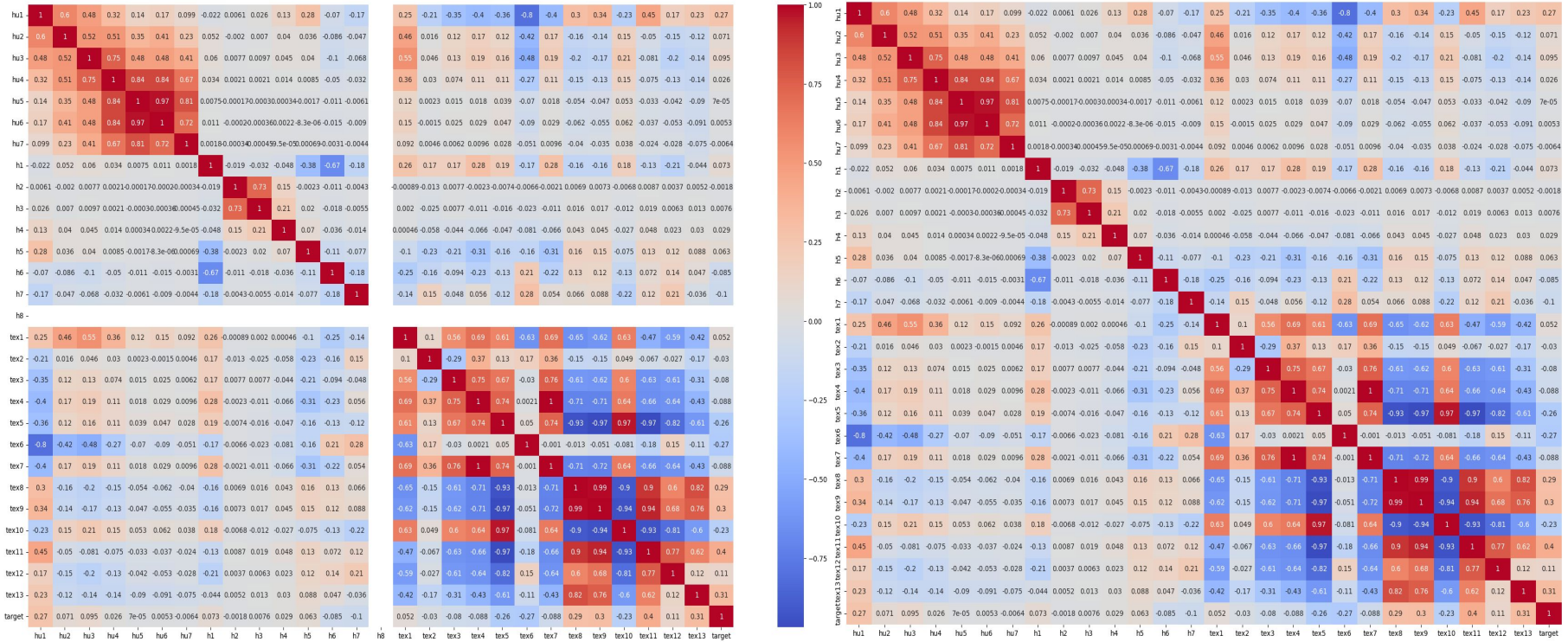
Preprocessing the dataset

- Unstructured pixel data in image patches is inadequate for direct classification and necessitates a suitable representation resilient to translation, rotation, and intensity variations.
- The primary challenge in Plasmodium detection lies in the geometry of objects within input segments, requiring a representation that is size-, translation- and rotation-invariant.
- Three edge detection filters—Canny, Sobel, and Scharr—were utilized to enhance the dataset by highlighting purple-stained regions in Parasitized Cell Images.
- Image preprocessing involved Histogram Equalization, but simplistic approaches generated image noise due to global and local contrast considerations.
- Adaptive Histogram Equalization, known as CLAHE (Contrast Limited Adaptive Histogram Equalization), was employed to limit contrast and perform precise histogram equalisation on small images and tiles.

Visualizing the dataset



Visualizing the dataset



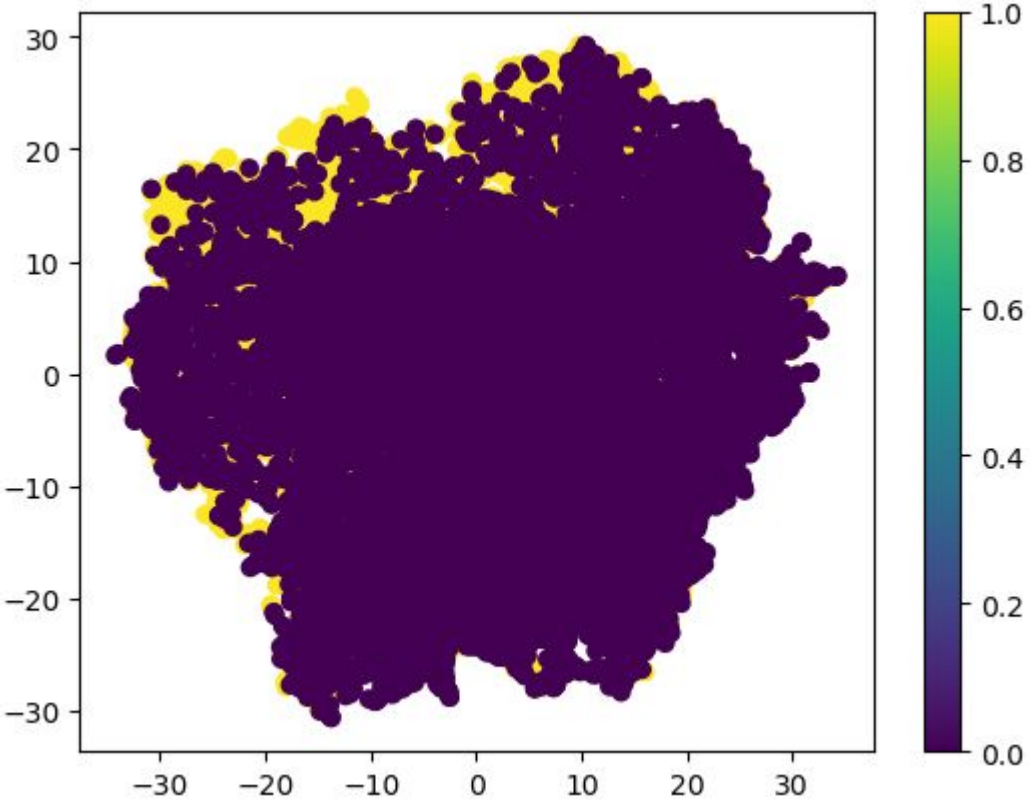
This heatmap shows correlation between different features. Feature “h8” isn’t correlated with any of the other features so, we decided to drop it.

Visualizing the dataset

This is the TSNE plot with a color gradient indicating the clustering and density of the data points. Here are a few interpretations based on the visual provided:

Purple Coloured points: Most of the data points that are visible in the plot are concentrated in the dark purple area. It represents a significant cluster of data and contains a substantial portion of data.

Yellow coloured points: The yellow color points also represents a distinct cluster of data points. However, due to the possibility of overlapping with the purple points, the yellow cluster may be partially obscured or hidden behind the purple cluster.



Visualizing the dataset

- Graphs at : LINK
- Pair-plots are also a great way to immediately see the correlations between all variables.
- Light shades represents positive correlation while Darker shades represents negative correlation in heatmap.
- From the pair plots we can observe that among all the factors leading to a malaria detection, the **Texture features** play a more impactful role:

Methodology

- We have utilized numerous filters, including edge detection algorithms Canny, Sobel, and Scharr, as well as Adaptive Histogram Equalization, also known as CLAHE
- We are splitting the dataset into a training and testing set using an 75:25 random split.After the division of our dataset, we chose supervised learning models to train and test on the dataset.
- We used local feature extraction techniques such as SIFT (Scale-Invariant Feature Transform) and KAZE in addition to Global feature extraction.
- These are following models that we implemented for training the data in addition to the edge detection algorithms:

  Logistic regression
  Random forest
  K nearest neighbours
  Ada-boosting
  CNN
  SVM
  Decision Tree
  Naive Bayes
  Grid search

Methodology

- We then used ensembling techniques, such as Bagging and Boosting, to determine if we could improve the accuracy.
- We used the Random Forest Classifier for Bagging
- We conducted Boosting with both Logistic Regression and the Decision Tree using Adaboost.
- After applying traditional models such as SVM, Random Forest, AdaBoost, Decision tree, KNN, Naive bayes ,etc., we attained an accuracy of 91.9 percent.
- We decided to use Deep Learning models to create a robust system that exploits the properties of images effectively. (CNN)

Result and Analysis

Model Performance on Raw data:

Model	Accuracy	Precision	Recall	F1
Logistic	0.864	0.878	0.848	0.863
SVM	0.872	0.881	0.864	0.873
DT	0.726	0.706	0.787	0.744
RF	0.807	0.776	0.867	0.819
KNN	0.739	0.741	0.745	0.743
Ada Boost	0.726	0.706	0.787	0.744

Model Performance after Global Feature Extraction

Methods	Logistic	SVM	DT	RF	KNN	NB	AB
raw	0.864	0.872	0.694	0.801	0.739	0.494	0.772
canny	0.839	0.863	0.777	0.859	0.830	0.731	0.837
clahe	0.827	0.852	0.590	0.750	0.797	0.672	0.801
scharr	0.894	0.894	0.785	0.876	0.851	0.762	0.859
sobel	0.913	0.909	0.792	0.884	0.853	0.784	0.862

Model Performance after Local Feature Extraction

Feature	Logistic	SVM	RF
SIFT	0.894	0.903	0.906
KAZE	0.668	0.693	0.670

Result and Analysis

	Model	Logistic Accuracy	Logistic Precision	Logistic Recall	Logistic f1
0	raw	0.863570	0.877600	0.848363	0.862734
1	canny	0.839042	0.823915	0.866743	0.844787
2	clahe	0.826560	0.847674	0.800689	0.823512
3	scharr	0.894340	0.910310	0.877369	0.893536
4	sobel	0.912627	0.928571	0.896037	0.912014

	Model	SVM Accuracy	SVM Precision	SVM Recall	SVM f1
0	raw	0.872424	0.881113	0.864159	0.872553
1	canny	0.862700	0.862286	0.866743	0.864509
2	clahe	0.852395	0.855085	0.852384	0.853732
3	scharr	0.893759	0.907528	0.879380	0.893232
4	sobel	0.909144	0.931942	0.884836	0.907778

	Model	RF Accuracy	RF Precision	RF Recall	RF f1
0	raw	0.806531	0.776150	0.867318	0.819205
1	canny	0.858781	0.856089	0.866169	0.861099
2	clahe	0.754717	0.716216	0.852384	0.778390
3	scharr	0.877358	0.880739	0.875933	0.878330
4	sobel	0.884906	0.891868	0.878805	0.885289

Result and Analysis

Local feature extraction techniques to the filtered images:

Feature	Logistic	SVM	RF
canny	0.820	0.817	0.808
clahe	0.869	0.881	0.879
scharr	0.811	0.817	0.807
sobel	0.798	0.805	0.801

Fig. 5. Filtered Images with Sift

Feature	Logistic	SVM	RF
canny	0.746	0.749	0.745
clahe	0.784	0.797	0.795
scharr	0.690	0.697	0.691
sobel	0.686	0.691	0.689

Fig. 6. Filtered Images with Kaze

Grid Search on SVM:

	Accuracy	Precision	Recall	F1
SIFT	0.902	0.898	0.906	0.902
canny	0.869	0.865	0.877	0.871
clahe	0.865	0.864	0.869	0.866
scharr	0.906	0.923	0.889	0.906
sobel	0.919	0.938	0.899	0.918

Fig. 7. Grid Search on SVM

Result and Analysis

Boosting and Bagging Results:

	Accuracy	Precision	Recall	F1
SIFT	0.904	0.899	0.910	0.905
scharr	0.877	0.881	0.874	0.878
sobel	0.885	0.891	0.881	0.886

Fig. 8. Bagging with Random Forest

	Accuracy	Precision	Recall	F1
SIFT	0.878	0.920	0.828	0.872
scharr	0.853	0.856	0.853	0.855
sobel	0.840	0.854	0.825	0.839

Fig. 9. Adaboost with Logistic Regression

	Accuracy	Precision	Recall	F1
SIFT	0.876	0.923	0.820	0.868
scharr	0.852	0.849	0.861	0.855
sobel	0.853	0.842	0.872	0.857

Fig. 10. Adaboost with Decision Tree

Result and Analysis

Performance of CNN Model

Training Performance:

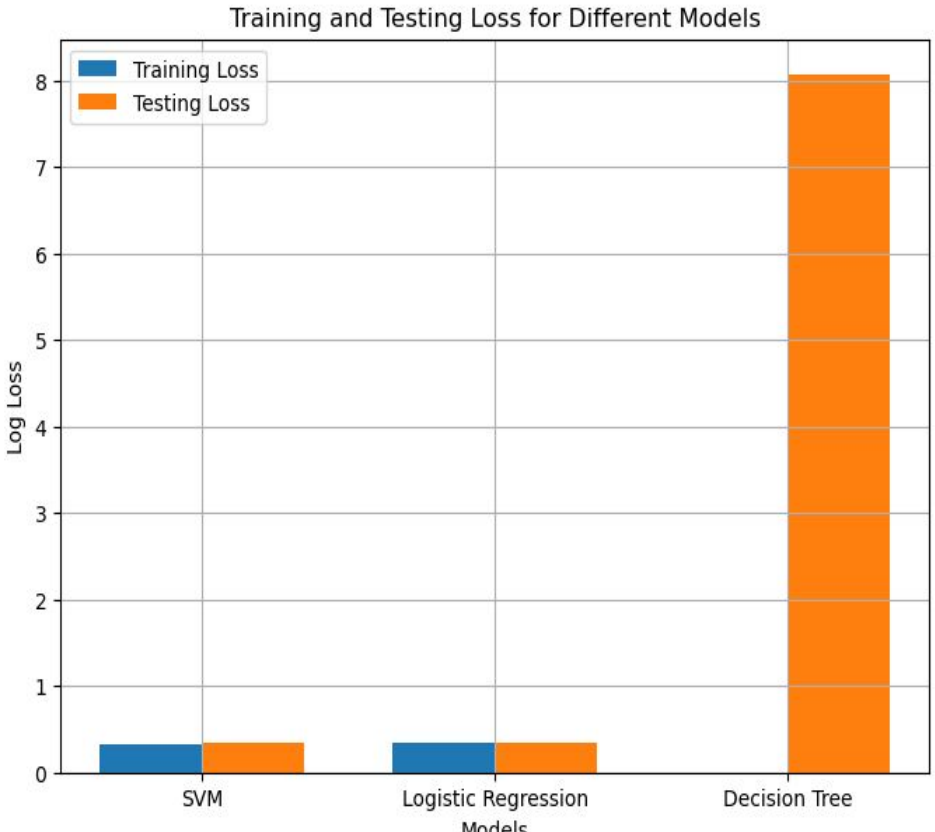
	Accuracy	Precision	Recall	F1
VGG11	0.998	0.997	0.999	0.998

Training Performance:

	Accuracy	Precision	Recall	F1
VGG11	0.940	0.923	0.961	0.941

Result and Analysis

The decision trees show signs of overfitting, so we decided to explore other models like SVM and logistic regression to achieve better generalization of performance and avoid the issues associated with overfitting. These models are chosen based on their strengths in handling various types of data and their ability to learn robust patterns for classification tasks



Conclusion

Based on the work done:

- We saw the importance of preprocessing steps and feature extraction techniques in enhancing model performance.
- Edge Detection Algorithm Sobel proved to be better than others (Scharr, CLAHE, canny).
- Logistic Regression demonstrated effectiveness, especially when used with appropriate preprocessing techniques.
- Support Vector Machine (SVM) emerged as a promising model among the baselines, achieving a good accuracy.
- Sift performed better than kaze in local feature extraction.
- Performance of Models was not as good as expected when local and global feature extraction were combined.
- Bagging and Boosting Techniques weren’t as useful.
- CNN model proved to be promising and gave an excellent accuracy of 94%.

Timeline

In terms of getting the work done, we adhered to the schedule and completed all of the tasks. Here is the work that we have completed:

1. Planning and Problem Definition
2. Data Gathering
3. Data Preprocessing (Canny, Sobel, Scharr, Clahe)
4. Data Visualization
5. Global Feature Extraction(Canny, Clahe, Scharr, and Sobel)
6. Model Selection
7. Model Training (SVM, Logistic Regression, Decision Tree, Random Forest, KNN, Naive Bayes, and AdaBoost)
8. Analysis of Model Performance
9. Local Feature Extraction (Sift and Kaze)
10. Model Performance on combination of Local and Global Features
11. Grid Search
12. Ensembling, Bagging and Boosting
13. CNN
14. Analysis, Evaluation and Documentation

Contribution

Although each member contributed equally, here are the individual contributions:

Manan garg: Data gathering, Data Preprocessing, Training Models, Result Analysis, Report Writing, Making Presentation.

Vivek jain: Data gathering, Data Preprocessing, Training Models, Result Analysis, Report Writing, Making Presentation.

Tushar: Data gathering, Data Visualization, Training Models, Result Analysis, Report Writing, Making Presentation.

References

https://www.kaggle.com/

https://www.nih.gov/

https://www.javatpoint.com/data-preprocessing-machine-learning

https://towardsdatascience.com/how-to-perform-exploratory-data-analysis-with-seaborn-97e3413e841d

https://www.ijrar.org/papers/IJRAR19SP014.pdf

https://www.ijert.org/malaria-detection-using-image-processing-and-machine-learning

https://www.researchgate.net/publication/322459559_Image_analysis_and_machine_learning_for_detecting_malaria

https://www.researchgate.net/publication/343346310_Detection_of_Malaria_using_Machine_Learning

THANK YOU

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**