**19CSE305 - MACHINE LEARNING**

**CLASSIFICATION OF DIABETIC RETINOPATHY**



**Group Details**

|  |  |  |
| --- | --- | --- |
| **S.No** | **Group Members** | **Roll Numbers** |
| 1. | Abineha P | CB.EN.U4CSE21201 |
| 2. | Apoorva S B | CB.EN.U4CSE21205 |
| 3. | Suruthi Lavanya A | CB.EN.U4CSE21262 |
| 4. | Tadepalli Sri Valli Ramya | CB.EN.U4CSE21263 |

**TABLE OF CONTENTS**

**Chapter 1:** Introduction.................................................................

**Chapter 2:** Block Diagram.................................................................

**Chapter 3:** Dataset Details...............................................................

**Chapter 4:** ClassificationReport.....................................................

**Chapter 4.1:** Performance evaluation....................................

**Chapter 4.2:** Inference of Result.......................................

**Chapter 5:** Conclusion...........................................................

**Chapter 6:** References..............................................................

**INTRODUCTION**

**Diabetic retinopathy** is a complication of diabetes that affects the eyes. It is considered a deadly eye condition as it can cause loss of vision and blindness among people who are diabetic. The high blood sugar levels cause considerable damage to the blood vessels in the retina and these blood vessels in the eye begin to leak fluid causing the macula to swell or thicken, preventing blood from passing through. Sometimes, there is an abnormal growth of new blood vessels on the retina. Diabetic retinopathy, when it is not detected and treated at the right time might lead to blindness eventually.

1. Provided with a large set of **high-resolution retina images** taken under a variety of imaging conditions, we create an automated analysis system capable of assigning a score based on the scale of severity of diabetes.
2. Predict whether a patient has diabetes, based on certain diagnostic measurements in the dataset. Several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

**MACHINE LEARNING ALGORITHMS:**

Supervised Learning:

1. Naïve bayes
2. LightGBM
3. Knn
4. Support vector classifier
5. Decision tree classifier
6. Random forest classifier

Unsupervised Learning:

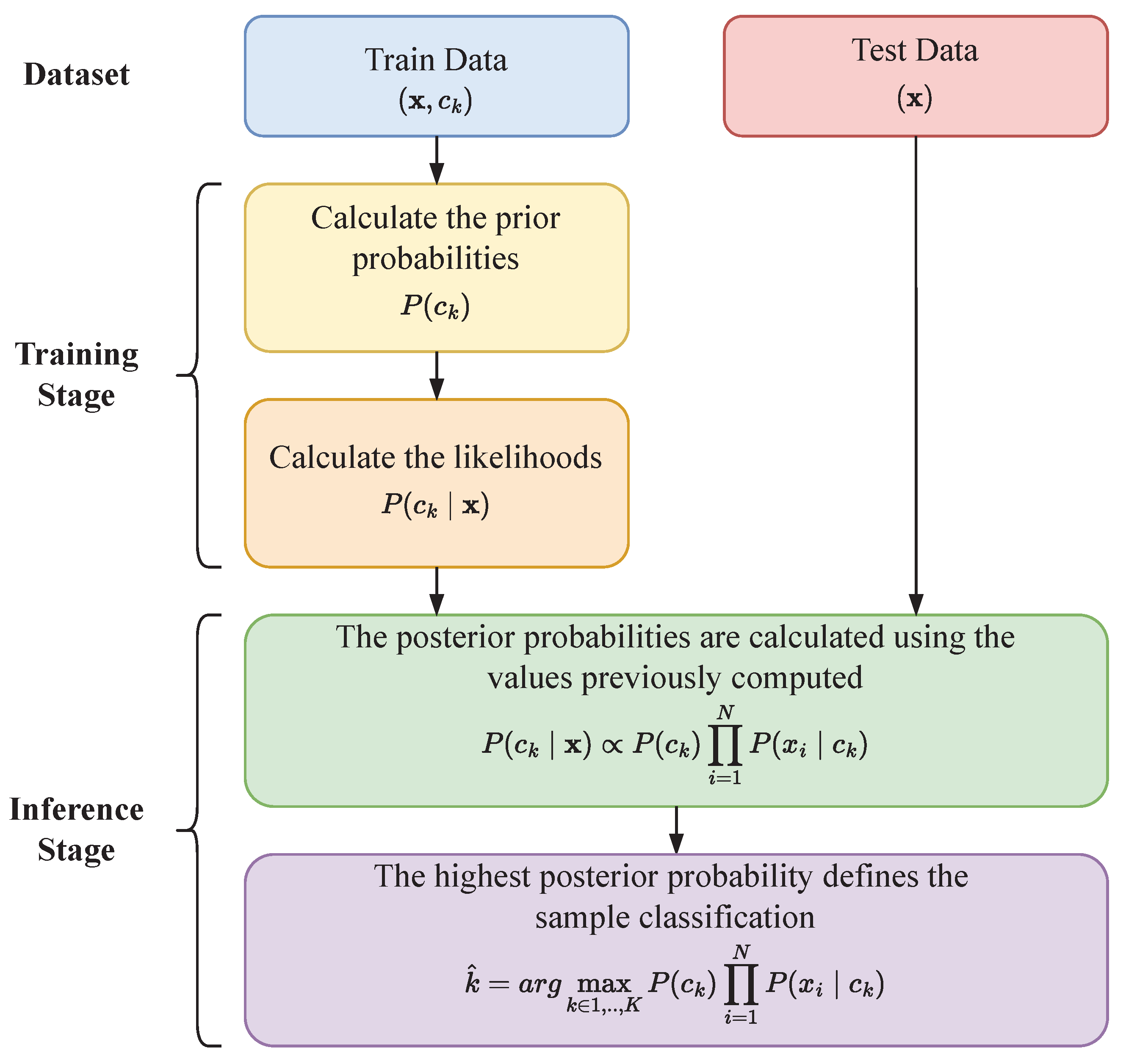
1. K means
2. Hierarchical Clustering (Ward’s Method)

**DEEP LEARNING APPROACH:**

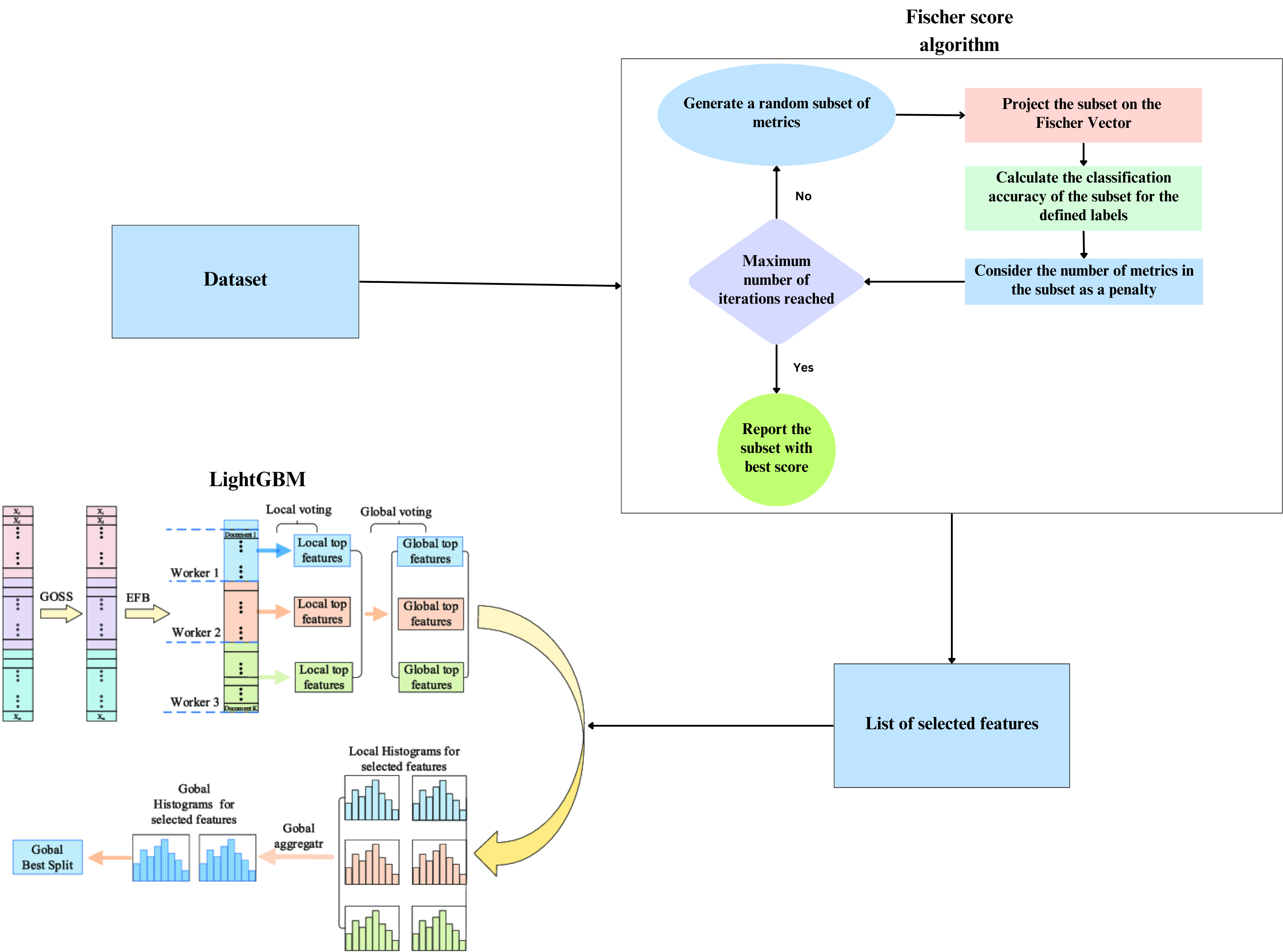
1. EfficientNet Convolutional Neural Network (EfficientNet CNN)

**BLOCK DIAGRAM**

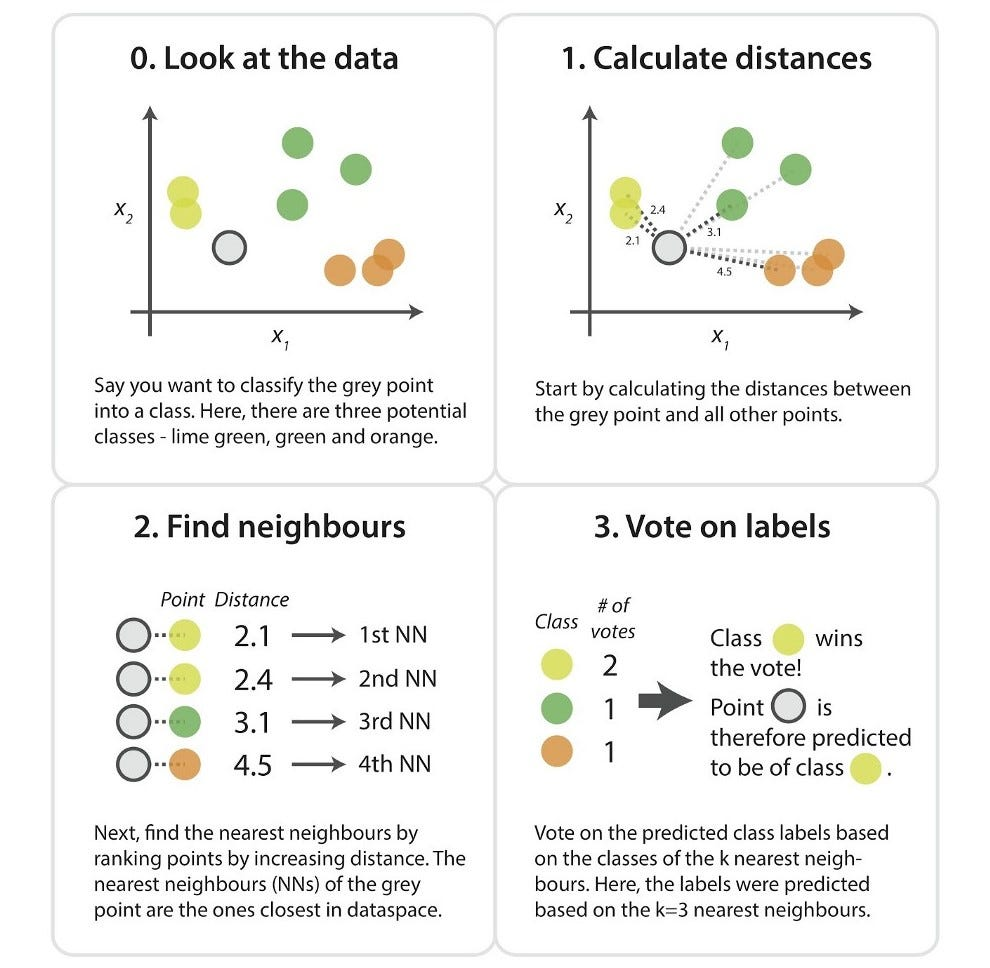
1. Naive bayes



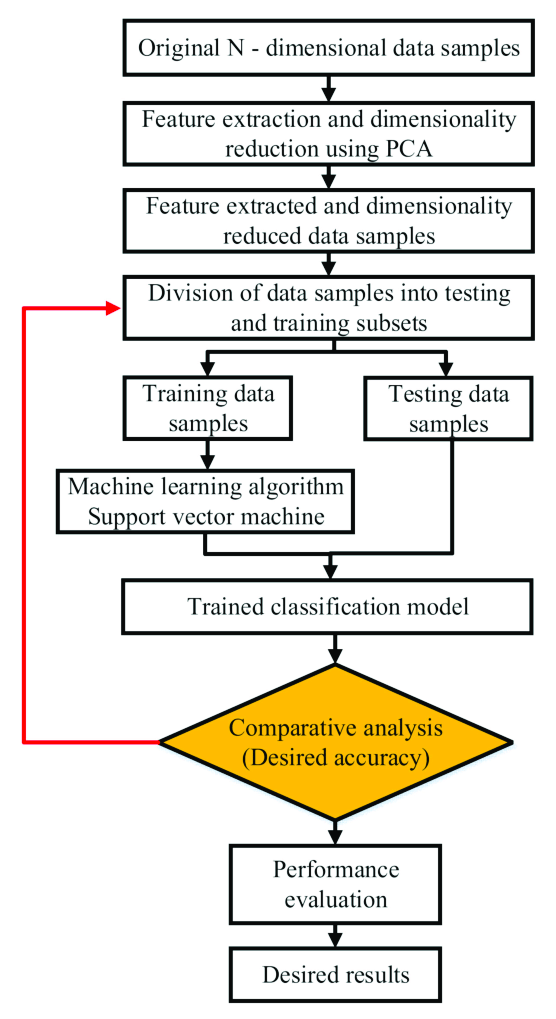
1. LightGBM



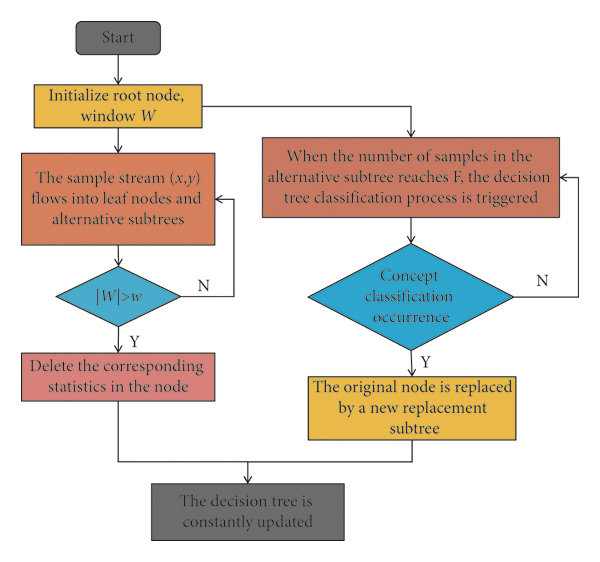
1. KNN Classification



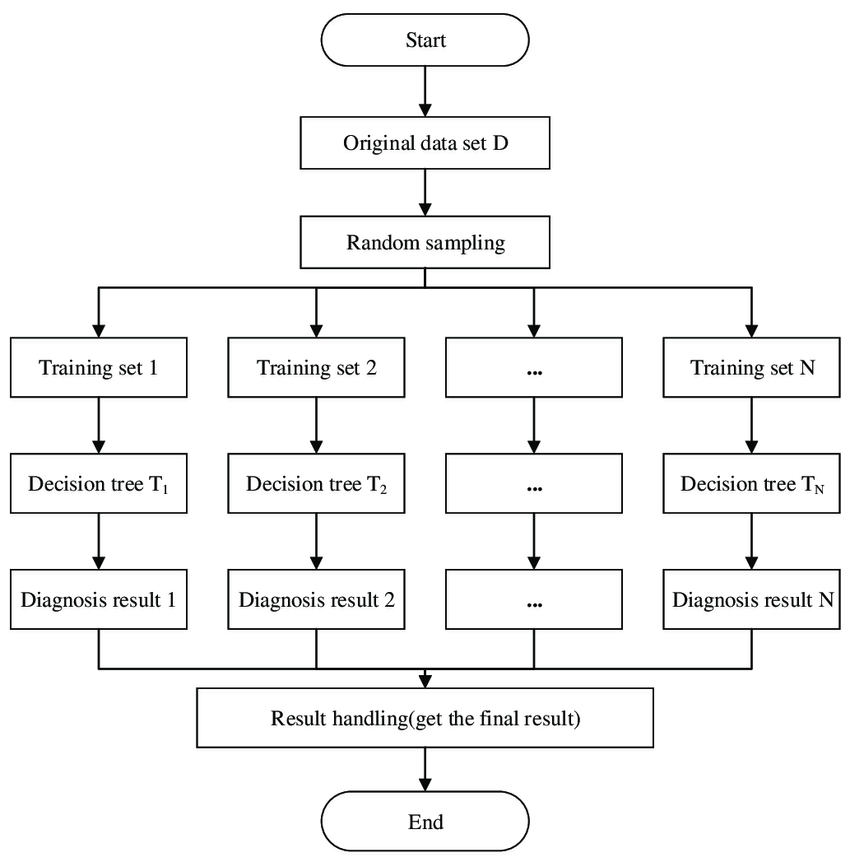
1. Support Vector Classification



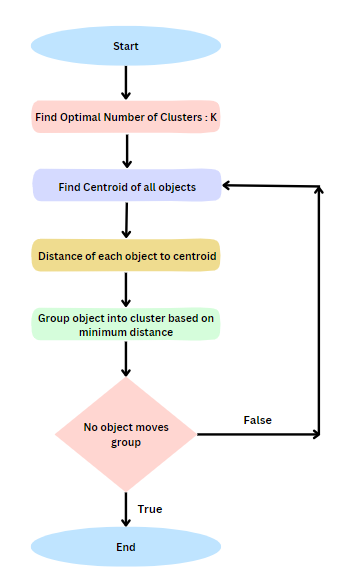
1. Decision tree classifier



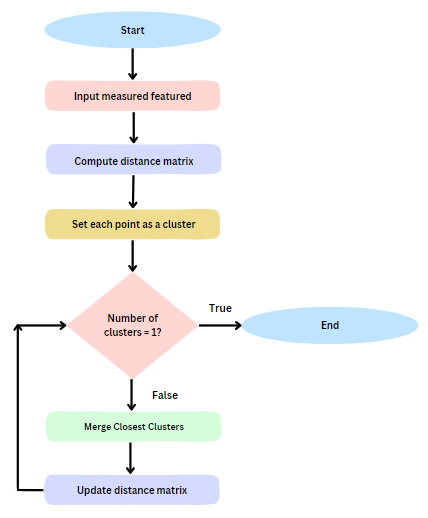
1. Random forest classifier



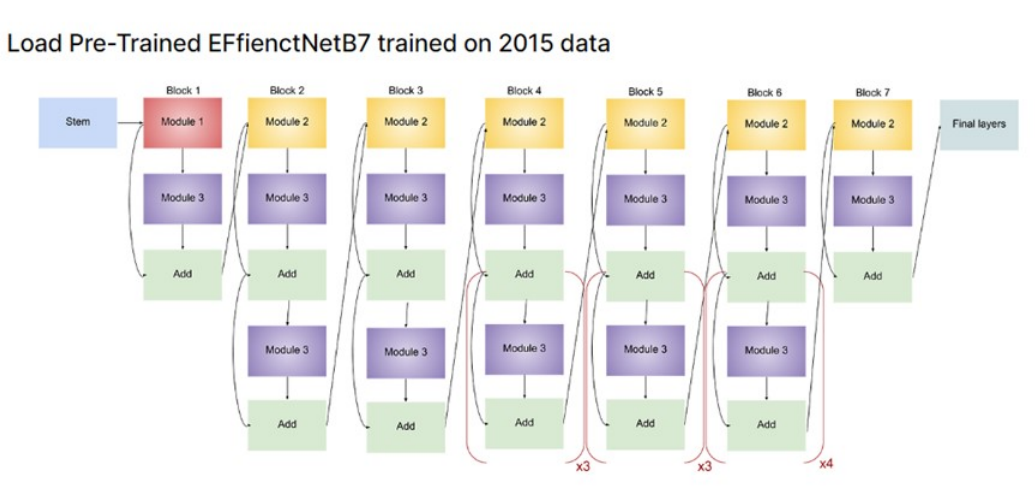
1. K – Means Clustering

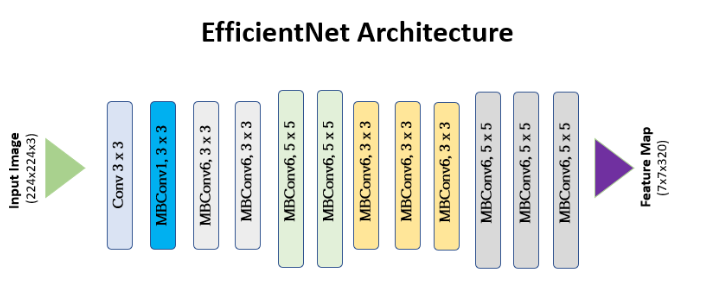


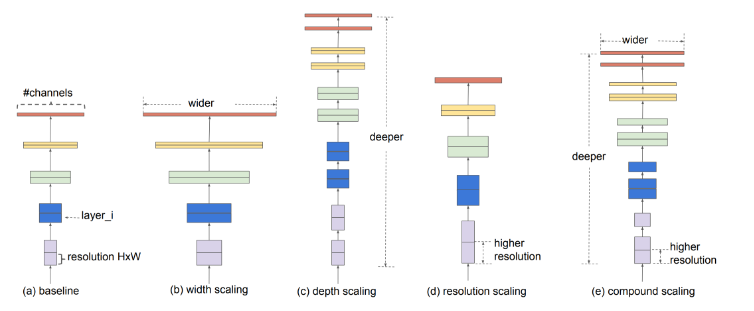
1. Hierarchical Clustering (Ward’s method)



1. EfficientNet







**DATASETS**

* **DATASET -** [[1]:](https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-%20using-machine-learning)

The **2015** dataset is a “Kaggle” dataset which is a large set of high-resolution retina images taken under a variety of imaging conditions. This dataset provides a left and right field for every subject. Images are labeled with a subject ID as well as either left or right. For example, 1\_left.jpeg is the left eye of patient id 1

In the dataset, the presence of diabetic retinopathy in each image is rated by a clinician on a scale of 0 to 4, according to the following scale:

* 0 - No DR
* 1 – Mild
* 2 – Moderate
* 3 – Severe
* 4 - Proliferative DR

The images in the dataset come from different models and types of cameras, which can affect the visual appearance of left vs. right. Some images are shown as one would see the retina anatomically (macula on the left, optic nerve on the right for the right eye). Others are shown as one would see through a microscope condensing lens (i.e., inverted, as one sees in a typical live eye exam). There are two ways to tell if an image is inverted:

* It is inverted if the macula (the small dark central area) is slightly higher than the midline through the optic nerve. If the macula is lower than the midline of the optic nerve, it is not inverted.
* If there is a notch on the side of the image (square, triangle, or circle) then it is not inverted. If there is no notch, it is inverted.

**Number of images:** 35,126

* **DATASET - [**[2]:](https://www.kaggle.com/c/aptos2019-blindness-detection/data)

The **2019** dataset is also a “Kaggle” dataset that contains a large set of retina images taken using fundus photography under a variety of imaging conditions. In the dataset, the presence of diabetic retinopathy in each image is rated by a clinician in the same order as that in the 2015 dataset. Transfer learning is applied to this dataset with the Knowledge Base acquired from the 2015 dataset to obtain more accurate and precise results.

Like any real-world data set, there exists noise in both the images and labels. The images may contain artifacts, be out of focus, underexposed, or overexposed. These images were gathered from multiple clinics using a variety of cameras over an extended period, which will introduce further variation.

**Number of training images:** 3,662

**Number of testing images:** 1,928

* **DATASET-** [[3]:](https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/input)

The dataset is a numerical dataset that consists of several medical predictor (independent) variables and one target (dependent) variable, Outcome. Independent variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

* + **Pregnancies:** Number of times pregnant
  + **Glucose:** Plasma glucose concentration 2 hours in an oral glucose tolerance test
  + **BloodPressure:** Diastolic blood pressure (mm Hg)
  + **SkinThickness:** Triceps skin fold thickness (mm)
  + **Insulin:** 2-Hour serum insulin (mu U/ml)
  + **BMI:** Body mass index (weight in kg/ (height in m) ^2)
  + **DiabetesPedigreeFunction:** Diabetes pedigree function
  + **Age:** Age (years)
  + **Outcome:** Class variable (0 or 1)

**Number of Observation Units:** 768

**Variable Number:** 9

**PERFORMANCE**

**Supervised Learning:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.No** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1\_Score** |
| 1. | Naive Bayes | 0.79 | 0.76 | 0.74 | 0.75 |
| 2. | LightGBM | 0.81 | 0.77 | 0.77 | 0.77 |
| 3. | KNN | 0.87 | 0.86 | 0.85 | 0.85 |
| 4. | SVM Classifier | 0.87 | 0.86 | 0.85 | 0.85 |
| 5. | Decision Tree Classifier | 0.87 | 0.85 | 0.85 | 0.85 |
| 6. | Random Forest Classifier | 0.91 | 0.89 | 0.92 | 0.90 |

**Unsupervised Learning:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.No** | **Method** | **Silhouette Score** | **Davies – Bouldin Score** |
| 1. | K Means Clustering | -0.04866 | 5.67022 |
| 2. | Hierarchical Clustering (Ward’s) | 0.15675 | 2.36392 |

**Neural Network:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Method** | **Accuracy** | **Precision** | **Recall** | **F1\_Score** | **OOF loss** | **OOF kappa** |
| 1. | 2015 dataset | 0.58 | 0.53 | 0.58 | 0.54 | 1.0675 | 0.6855 |
| 2. | 2019 dataset | 0.73 | 0.58 | 0.73 | 0.64 | 0.8921 | 0.7696 |

**CONCLUSION**

1. Apart from imagery data, numerical data is also processed across various algorithms namely Naive Bayes, [[3]](https://www.kaggle.com/code/ahmetcankaraolan/diabetes-prediction-using-machine-learning/input) LightGBM, KNN, and K-means clustering, and attained the maximum possible accuracy with every model to enable one to understand the importance of the proposed EfficientNet algorithm for the imagery data. As a result, it is proven that the proposed algorithm performs better than other existing algorithms in real-world scenarios. The algorithm with highest accuracy and precision is **Random Forest Classifier**.

2. **K - Means and Hierarchical clustering** do not really work why?

* **Overlapping Clusters:** The natural separation between individuals with diabetes and those without it is complex, leading to overlap between the clusters. Both hierarchical clustering and K-means assume distinct clusters, and since the data does not exhibit clear separation, the algorithms do not perform well.
* **Non – Linear relationship between features and outcome**: The relationship between the features and outcome does not have a direct linear relationship. Clustering methods are sensitive to the linearity of relationship between features and outcome and hence these two methods do not perform well.

3. The proposed **EfficientNet** algorithm along with an optimizer like Adam optimizer is aimed at enhancing the training efficiency and effectiveness of the proposed neural network. The training efficiency and effectiveness are improved with faster training time, improved handling of vanishing or exploding gradients, and potentially better generalization on unseen data.

* Performance: Due to its depth and capacity, EfficientNetB7 can achieve high accuracy and robust performance on a wide range of computer vision tasks, including image classification, object detection, and segmentation.
* Transfer Learning: EfficientNetB7 is often used as a powerful feature extractor for transfer learning. Researchers and practitioners fine-tune this model on specific tasks, leveraging the pre-trained weights and representations learned from large-scale datasets.
* Reason**:** It can be applied to various computer vision tasks, including medical image analysis, autonomous driving, and natural image classification. EfficientNetB7 requires loading a pre-trained version of the model and fine-tuning it for a specific task. The model’s strong feature extraction capabilities help in adapting itself to a unique dataset and problem.

**REFERENCES**

**Datasets:**

1. <https://www.kaggle.com/c/diabetic-retinopathy-detection>
2. <https://www.kaggle.com/c/aptos2019-blindness-detection/data>
3. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database>

**Base Papers:**

1. <https://www.sciencedirect.com/science/article/pii/S1877050920300557>
2. <https://ieeexplore.ieee.org/document/9175664>