A black background with white text

Description automatically generated with low confidence

MSc Data Science

Individual Project Module

7PAM2002-0901-2024

Department of Physics, Astronomy and Mathematics

**MSc Data Science Project Handbook**

**2024-25**

**Semester A**

**MSc Data Science Project Module Leader**

**Carolyn Devereux**

[c.devereux@herts.ac.uk](mailto:c.devereux@herts.ac.uk)

**University of Hertfordshire**

**College Lane, Hatfield,**

**Herts AL10 9AB, United Kingdom**

Predicting Diabetic Retinopathy Using Machine Learning on the Diabetic Retinopathy Debrecen Dataset

**GitHub Link**:  
<https://github.com/Warraich28/diabetic_project>

Student Name: Muhammad Noman Naeem  
Supervisor Name: Hassan al madfai

MSc Program: MSc Data Science, University of Hertfordshire

**Word Count**: 8111 words

**Acknowledgements**

I wish to convey my heartfelt gratitude to my supervisor for their pivotal role as a source of guidance, encouragement, and constructive feedback throughout this work. I would like to express my gratitude to the Data Science Department of the University of Hertfordshire for providing the academic resources and insights necessary to complete this work. Tremendous gratitude to family and friends for their support, particularly during the most challenging aspects of my endeavors.

**Abstract**

This research investigates the ability of machine learning models in predicting diabetic retinopathy using the Diabetic Retinopathy Debrecen dataset, which comprises 1151 records with features extracted from retinal images. This study aimed to evaluate the efficacy of Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, and an ensemble classifier combining Logistic Regression and Random Forest models. Thus, the analysts incorporated data preprocessing, feature selection utilizing Random Forest feature importance, and hyperparameter optimization for rapid and precise predictions. Logistic Regression emerged as the best-performing model with an accuracy of 75.86%, followed closely by the Hybrid Voting Classifier with 75.00%. Random Forest demonstrated effective predictive capabilities; however, SVM and Naïve Bayes exhibited deficiencies in recall and overall accuracy. The results indicate that machine learning models can be trained to automatically identify diabetic retinopathy; however, models that are accurate, efficient, and easily interpretable, specifically Logistic Regression and the Hybrid Voting Classifier, are optimal for practical applications in diabetic retinopathy detection. The research conclude that these models are effective for early diabetic retinopathy screening and vision loss prevention, and advocate for future research like larger datasets, advanced deep learning models, temporal data to enhance predictive accuracy and generalization.

**Table of Contents**

[Chapter 1: Introduction 6](#_Toc184815437)

[1.1 Aims and Objectives: 7](#_Toc184815438)

[Chapter 2: Background and Literature Review 8](#_Toc184815439)

[Chapter 3: Dataset 11](#_Toc184815440)

[3.1 Exploratory Data Analysis (EDA) 11](#_Toc184815441)

[3.2 Data Preprocessing 13](#_Toc184815442)

[Chapter 4: Methodology 14](#_Toc184815443)

[4.1 Model Selection 14](#_Toc184815444)

[4.1.1 Support Vector Machine 14](#_Toc184815445)

[4.1.2 Naïve Bayes 15](#_Toc184815446)

[4.1.3 Random Forest 15](#_Toc184815447)

[4.1.4 Logistic Regression 15](#_Toc184815448)

[4.2 Model Evaluation Metrics 16](#_Toc184815449)

[4.3 Hyperparameter Tuning 17](#_Toc184815450)

[Chapter 5: Results 18](#_Toc184815451)

[Chapter 6: Analysis and Discussion 21](#_Toc184815452)

[Conclusion 24](#_Toc184815453)

[Bibliography 26](#_Toc184815454)

# Chapter 1: Introduction

Diabetic retinopathy (DR) (Fong, 2003) is one of the most severe complications of diabetes and is a main reason of blindness in working-age adults globally. Early detection and accurate prediction of DR are crucial to prevent irreversible damage to vision. Diabetic retinopathy develops silently, with minimal symptoms in the early stages, but can progress to more severe forms that threaten vision if left undiagnosed or untreated​ (Gadekallu T. R., 2023). Traditionally, DR is diagnosed using retinal imaging, where ophthalmologists analyze retinal photographs for abnormalities such as microaneurysms, hemorrhages, and neovascularization. In practice, the evaluation of these images can often be labour-intensive and contain human bias, particularly when used in large-scale studies. This has created a focus on building DL and ML models to detect DR and predict DR progression automatically​​ (Gadekallu T. R., 2020).

The aim of this work is to develop and compare machine learning models that predict the presence of DR using the Diabetic Retinopathy Debrecen dataset. This collection of records consists of 1151 records with retinal images and 19 features overall. In this paper, the used models provide a list of SVM, Naïve Bayes, random forest, and logistic regression to find out the most suitable one for predicting DR. Furthermore, the impact of different heuristics combined in a model shall be analysed to enhance the prediction ability of the model.

Using machine learning for image diagnosis has been relatively a fast-growing field in the last decade. Contrary to traditional programming, ML algorithms operate iteratively to discover patterns from the data and enhance prediction performance, which makes it possible to classify different diseases such as diabetic retinopathy. Retinal images are addressed by contemporary deep learning models as essential approaches for diagnosing DR and estimating the further disease evolution.For instance, DeepDR Plus was developed to predict the future progression of DR from retinal images and showcased powerful outcome in internal and external validation (Odeh, 2021). In another study, authors proposed a DR classification model by employing the DNN with GWO to classify DR based on extracted retinal image features​ Such developments envisage the possible role of machine learning in enhancing the effectiveness of DR diagnosis and eradicating the avoidable vision loss that burdens the health sectors.

This dissertation seeks to respond to the following research questions. This work investigates if machine learning is capable of accurately detecting the presence of DR only from features extracted from the retinal images themselves. Besides, it explores the predictive models such as Naïve Bayes, Support Vector Machine (SVM), Random Forests, and Logistic Regression and set up their comparative analysis regarding the accuracy of diagnosing diabetic retinopathy. Last and not least, the study looks at whether there is any improvement achieved when using composite models whereby multiple machine learning algorithms are used for the classification of diabetic retinopathy.

## 1.1 Aims and Objectives:

This study also seeks to develop an accurate, self-learning algorithm for screening retinal images in diabetic patients with diabetics retinopathy. In an effort to realize this goal the following objectives have been set: First, the Diabetic Retinopathy Debrecen dataset will be dealt firstly with methods of data preprocessing including handling missing values, normalization of the numerical features of a dataset, and performing feature engineering if needed for improving the efficiency of models. Subsequently, classification models such as SVM, Naïve Bayes, Random Forest, Logistic Regression will be excellent in predicting diabetic retinopathy (Sullivan, 2022). Furthermore, more intricate model that integrates two or more algorithms to arrive at a better prediction will be applied (Yang, 2020). The research will also entail model fine tuning known as hyperparameter tuning, through cross- validation techniques like grid search and random search. The Descriptive Analysis shall be done in order to generate hypotheses of the insights concerning the relationship between features in the set. Lastly, the merits of the models will be compared to such factors as precision, accuracy, recall, F1-measure, and AUC-ROC to and note the optimal algorithm.

# Chapter 2: Background and Literature Review

Diabetic retinopathy is one among the most important diabetic complications and the leading cause of blindness among people of working-age. A key challenge, in combating the progression to severe vision loss, is detecting DR in its early stage. Thanks to the use of the ML, new techniques have been introduced to teach models for detection and diagnosis of DR based on the retinal images that may help the ophthalmologists. Some techniques that have been applied in the enhancement of DR detection system are the application of deep learning model and dimension reduction. This paper aims at applying these ML methods to predict DR using the Diabetic Retinopathy Debrecen dataset.

Gadekallu (2020) put forward an accountable Machines learning model that includes principal component analysis and a firefly algorithm along with a deep neural network to identify diabetic retinopathy. Diabetic Retinopathy A Debrecen dataset was used in the study, and the preprocessed data was normalised using StandardScaler to reduce bias arising from scaling of the features. PCA was used to perform feature selection and dimensionality reduction on the samples, and then Firefly was used for optimisation of the feature selection for better results. The optimised dataset was then passed through a DNN model and returned an accuracy of 97%. This model fared better than the typical conventional machine learning algorithms such as Support Vector Machines (SVM) and Naïve Bayes. However, the training of deep neural networks with optimisation algorithms could be a drawback due to its computational intensity in practical applications, especially in applications with limited resources. However, the study did not consider cross-validation techniques, which could have supplemented the stability of the model outcomes. However, this paper is pertinent to the current project since it shows how dimensionality reduction approaches can improve one’s deep learning models for DR detection (Gadekallu T. R., 2020).

Gadekallu (2023) proposed the use of a deep learning method for identifying diabetic retinopathy from the fundus images of the eyes. The authors utilised a database containing 128,175 images in order to train a CNN that is capable of automatically identifying features of DR. The model described above obtained 90% sensitivity and 98% specificity, which proves that it can achieve the same results as ophthalmologists in cases of DR identification. One of the study’s main strengths was data coverage—one hundred instances were used to advance the generalisability of the model while reducing the probability of overfitting. Moreover, pre-built layers of the CNN model removing the need for feature extraction made it less time-consuming in diagnosing diseases. However, the study is focused on the fundus images only, while there is more information about DR available with other types of imaging like optical coherence tomography (OCT). In general, this paper is very useful for the project since it reveals the ability of deep learning models to automate DR identification, mainly when working with massive amounts of data (Gadekallu T. R., 2023).

Autonomous from all aforementioned, (Gundluru, 2022) presented a DL-based scheme to enhance the identification of diabetic retinopathy. The authors used a convolutional neural network (CNN) for the classification of retinal images based on stages of diabetic retinopathy. Rotation, flipping, and scaling were performed on images to increase the model’s robustness. Every performance of the model was considered using common parameters such as accuracy, sensitivity, and specificity. The CNN model has better accuracy with 97% against most traditional models that offered at most 85% accuracy. The paper also demonstrated that by integrating deep learning models with the data augmentation approach, small datasets could be managed, and the detection of diabetic retinopathy could be optimised. The results of this study show that DR can be detected using deep learning; data augmentation proves to be beneficial in improving model performance.

(Emon, 2021) compared various machine learning models to predict DR with the Diabetic Retinopathy Debrecen data set sourced from the UCI Machine Learning Repository. The work used several different algorithms, such as the Naïve Bayes algorithm, SMO, logistic regression, SGD, bagging classifiers, J48, decision trees, and random forest algorithms. Results in terms of accuracy, sensitivity, and ROC curves for each of the models were compared. When evaluating all the tested models, logistic regression showed the best performance with an accuracy of 75% out of all the other models. The paper concludes that traditional machine learning models like logistic regression are effective for DR prediction when applied to structured datasets such as the Debrecen dataset. However, it is important to note that in the course of the research, the possibility of employing deep learning mechanisms that can enhance the reliability of large data sets with features of images was not taken into account.

Based on an article by (Odeh, 2021) an ensemble learning framework that combines other ML algorithms to improve the diagnosis of diabetic retinopathy. For the study, the literature used the Messidor dataset, an archive of retinal fundus images, and feature selection methods like InfoGainEval and WrapperSubsetEval to create feature subsets. Others include random forests and neural networks; the support vector machine was included to enhance diagnosis, among other works in the ensemble learning model. When the authors retrained using the reduced set of features, however, the accuracy of the ensemble model stood at 71.5 percent. The authors stress that the combination of the ensemble learning with the feature selection illustrates an opportunity to improve the DR detection systems and decrease the model’s dimensionality.

Table 1 Litreature Review

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Methods** | **Dataset** | **Best Model Accuracy** |
| (Gadekallu T. R., 2020) | PCA, Firefly optimization, DNN | Diabetic Retinopathy Debrecen | 97% |
| (Gadekallu T. R., 2023) | CNN | Large retinal fundus image set | 90% Sensitivity, 98% Specificity |
| (Gundluru, 2022) | CNN with data augmentation | Public retinal fundus image set | 97% |
| (Emon, 2021) | Naive Bayes, SMO, Logistic Regression, SGD, Bagging, J48, RF | Diabetic Retinopathy Debrecen | 75% (Logistic Regression) |
| (Odeh, 2021) | Ensemble learning (Random Forest, SVM, Neural Networks) | Messidor dataset | 75.1% |
| - | SVM, Random Forest, Logistic Regression, Hybrid Models | Diabetic Retinopathy Debrecen | - |

From the analyzed papers, it is clear that implementing ML approaches, hence deep learning approaches, improves the detection and diagnosis of diabetic retinopathy. This was done by Gadekallu et al. (2020) in a method that incorporates dimensionality reduction techniques such as PCA combined with optimisation models to boost the classification results. Likewise, Gulshan et al. have used the method of convolutional neural networks to demonstrate that DR can be diagnosed automatically with increased sensitivity and specificity. Gundluru et al. (2022) also highlighted the same aspect but also how CNNs, if backed by data augmentation, impose a greater ability of generalisation of the model. However, Emon et al. (2021) and Odeh et al. (2021) used traditional machine learning and ensemble methods that were equally impressive in a standard structured dataset (Gadekallu T. R., 2020) (Gundluru, 2022) (Emon, 2021) (Odeh, 2021).

Therefore, this work will investigate using similar, complex methods such as the SVM and Random Forest as well as the logistic regression in some of the specific hybrid models in order to enhance the predictive accuracy on the Diabetic Retinopathy Debrecen data set. The project will aim to achieve similar or better accuracy to the previous studies to extend the literature on automated DR detection.

# Chapter 3: Dataset

The dataset utilized in this project is the Diabetic Retinopathy Debrecen dataset, sourced from the UCI Machine Learning Repository (Antal, 2014). This dataset is publicly accessible for research purposes and was originally collected in 2014 from the Messidor image set, intended to support research on diabetic retinopathy (DR) detection. This dataset has 1151 instances and 19 attributes, which are obtained from the evaluation of retinal images: integer, binary, and continuous. The main goal is to predict the presence of DR, with the target label being binary (1 for DR positive and 0 for DR negative). This dataset was selected because it relates to the topic of interest, that is, to employ machine learning methods to estimate the likelihood of developing DR. Due to its well-structured features and its public accessibility, it can be a reference for benchmarking of ML models on DR classification tasks.

The dataset used in this research is in CSV format, with a size of approximately 300KB. It includes features that represent the retinopathic appearance of the fundus, such as exudates and optic disc size, along with a target category indicating the diabetic retinopathy (DR) status. The data comes from the Messidor dataset in terms of retinal image features and was provided by researchers to the UCI (Antal, 2014) to facilitate the development of machine learning solutions for detecting diabetic retinopathy. It is anonymized to be ethical, and the kind of images used here are effective in model training in this project.

## 3.1 Exploratory Data Analysis (EDA)

The histogram of the class distribution that I have included in the following figure indicates almost equal numbers of DR positive and DR negative cases. Such balance is useful when training the machine learning model to avoid cases where the model will tend to favor one class over the other.

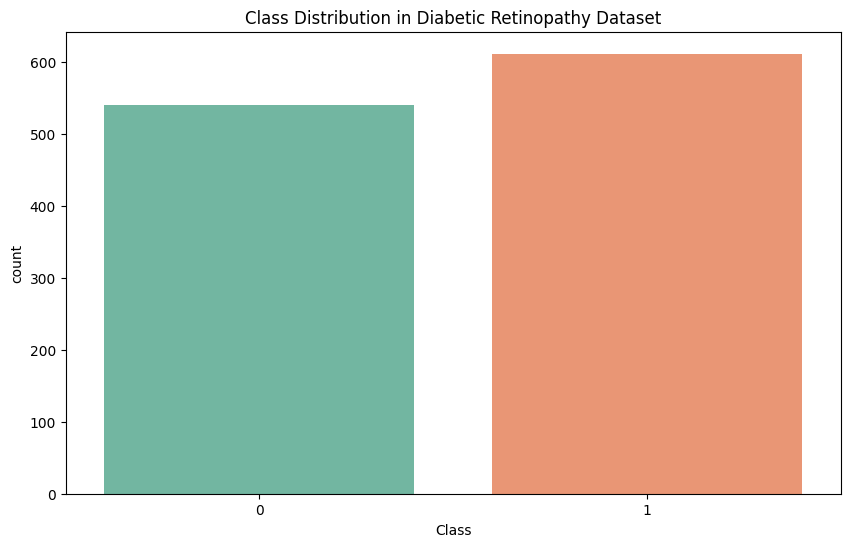


Figure Class Distribution Retinopathy Dataset

A heatmap of correlations (Tiessen, 2017) among features (attached image) highlights relationships between certain attributes, particularly among exudate-related features (e.g., exudate1, exudate2) and optic disc measurements like opticdisc\_diameter and macula\_opticdisc\_distance. Strong correlations suggest that some features may be redundant or dependent on others, which is crucial information for dimensionality reduction. For instance, highly correlated features might be combined or one of them might be removed to simplify the model, reduce overfitting, and improve computational efficiency.

Using a Random Forest classifier, feature importance analysis (see bar plot) identifies the most influential features, such as ma1, exudate1, and exudate2. These features have a higher impact on predicting DR and will guide the feature selection process (Menze, 2009), enabling the use of only the most relevant variables. This step helps streamline the model, focusing it on features with the greatest predictive power, which can improve model performance and reduce computational complexity.

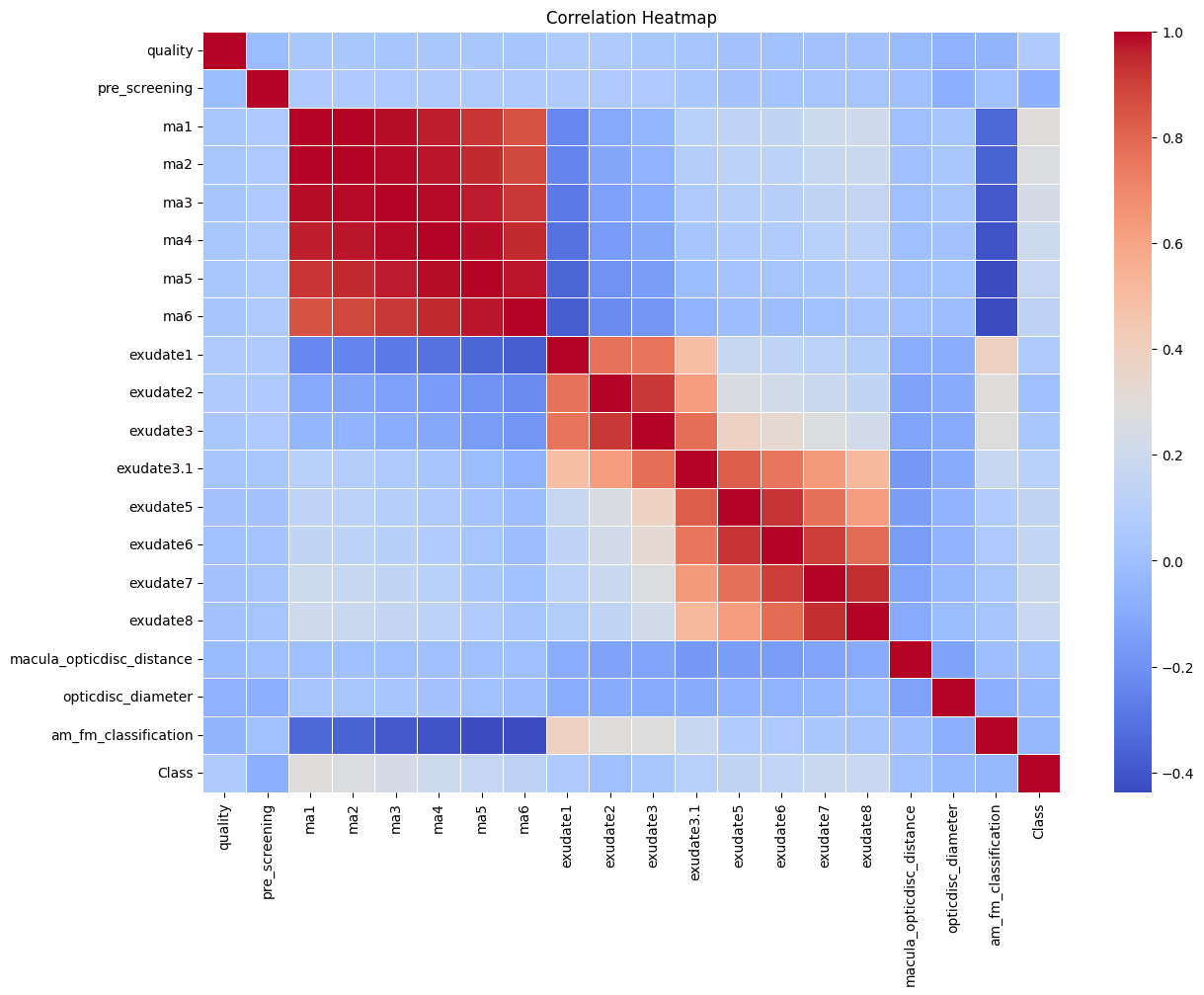


Figure Correlation Heatmap

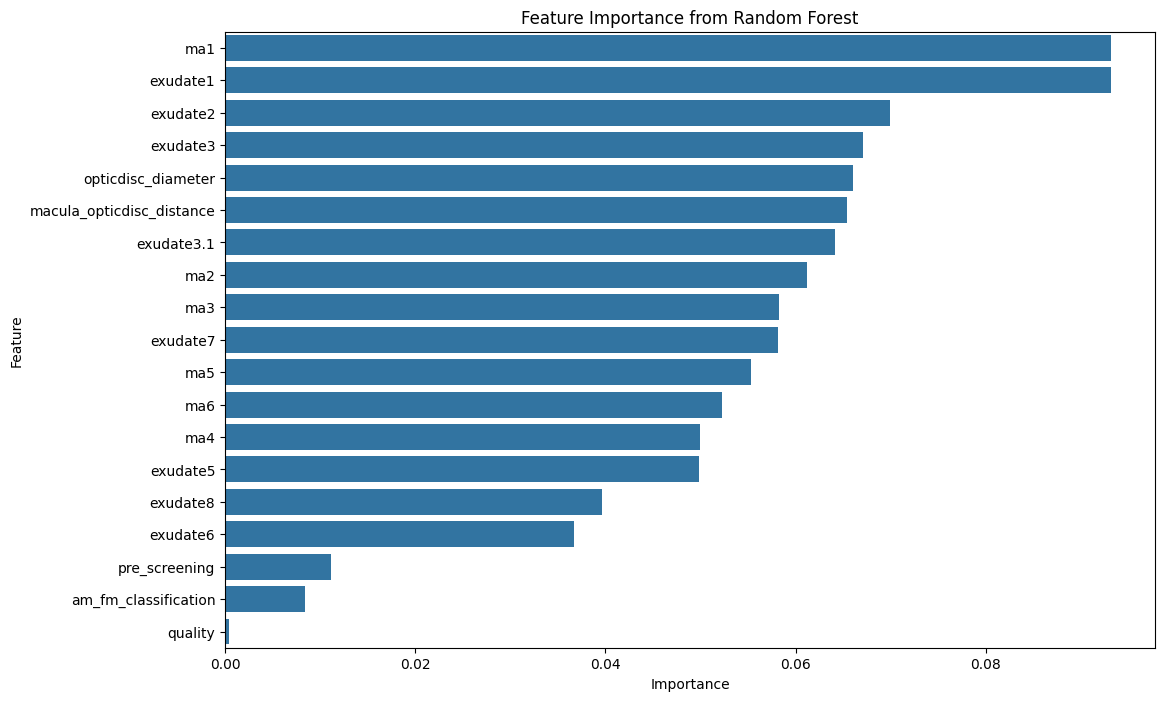


Figure Feature Importance using Random Forest

## 3.2 Data Preprocessing

In order to clean the data for analysis, some modifications were made as follows in order to enhance the reliability of the data. Initially, the presence of missing values in the set was examined; the presence of entire records that lack the critical information necessary for making predictions is inadmissible when creating a machine learning model. Actually, the evaluation of missing values was one of the most important tasks in maintaining data credibility. Subsequently, normalization (Çetin, 2022) was done to the continuous measures including exudate measurements and distances associated with the optic disk using StandardScaler. This process makes features to be in a standard range, which may benefit the machine learning algorithms since they are affected by difference in scale of features. Normalization reduces skew in features that have larger scales, and assures higher quality and stability in the model. Lastly, the dataset was split into training and testing sets in an 80:20 ratio. It also affords us the opportunity to estimate their performance on new data since it splits the modeling process into train and test data set, which reduces over-fitting encountered during training.

These and all the similar EDA findings (Camizuli, 2018) and preprocessing steps are essential for the correct setting of the model and its training and testing phase. It provides an idea about the structure and reduces the complexities involved in feature extraction for the model, and allows to input only the suitable and cleaned data for the model. It is considered to improve this capacity with the diabetic retinopathy issue in this venture through the application of this exhaustive method.

# ****Chapter 4: Methodology****

The main goal of this work was to attempt to classify DR using different AI algorithms on the given dataset, ‘Diabetic Retinopathy Debrecen’. The next sub-section, called methodology, will explain the method followed systematically, which includes feature selection, model selection, training, validation, and tuning. Every single level was planned precisely in order to maximize the correctness and stability of diabetic retinopathy prediction. Earlier on, feature selection was one of the main steps that were involved in the preparation of the dataset for modeling. To maximize the models’ efficiency and performance and, at the same time, minimize the computational load, I concentrated on the main features only. According to the Random Forest classifier, I obtained the feature importance analysis to know which variables had the highest impact on the model. Having set over some importance weights, I was able to afford to retain only the most important features and discard the nuisance features, which reduced variability. This process not only helped the model to focus on relevant data but also accelerated the training process and increased the accuracy we are achieving by minimizing the amount of data. Eventually, feature selection allowed the models to concentrate on the main determining factors of diabetic retinopathy, which are more accurate.

## 4.1 Model Selection

In order to have a full picture and compare the results, I chose and applied several machine learning models. The models selected for this study were Support Vector Machine, Naïve Bayes, Random Forest, and Logistics Regression. All those models were chosen because they have their own specific strengths, and therefore they were suitable to contribute something different in the classification of diabetic retinopathy.

### 4.1.1 Support Vector Machine

SVM is well suited for the task because it is proven exceptionally accurate for binary classification tasks, especially where the classes are convex and include a large margin (Vishwanathan, 2002). It improves the separability of the diabetic retinopathy data by transforming the inputs into a hyper plane feature space through kernel functions. Further, the classification probabilities of the model were developed using setting probability-based SVM in order to allow more flexible decision boundaries as depending on the outcome of the model and the peculiarities of the medical diagnostic task. SVM was looked to perform better than other models which made it very useful in generalizing the search acrossdifferent datasets as well as preventing the common problem of overfitting of models (Cortes, 1995).

Where w is the weight vector, b is the bias, ​ is the label (+1 or −1) for the data point ​.

### 4.1.2 Naïve Bayes

The first algorithm adopted from Webb (2010) was Naïve Bayes since it is a simple and fairly computationally intensive baseline. Despite the assumption that features should be independent and identically distributed which is not always true a high-dimensional dataset has been found to work unexpectedly well with this model. In addition, with Naïve Bayes it was possible to obtain meaningful data about the ability of the chosen features in the data set within a short time. Due to its low computational complexity Webb compared it with higher alternative models whose calculation was computationally intensive (Webb, 2010) (McCallum, 1998).

P(y ∣ x) is the posterior probability of class y given features x

P(y) is the prior probability of class y

is the likelihood of feature ​ given class y

P(x) is the probability of x

### 4.1.3 Random Forest

Mainly, Random Forest was chosen for this project because of its high accuracy and because of its stability within the model itself. This algorithm operates in a way that first averages many decision trees, of which each is trained on one random set of data. This approach helps to minimize variance and thereby reduces the danger of over-fit of the model. Random forest is more suitable for unbalanced data because of its capacity to work with high interactions of features as perceived by Rigatti (2017). Moreover, the feature importance analysis embedded in the model introduced a wealth of understanding into which feature specializes in prediction, which testified to its appropriateness to this task. This capability ensured compatibility of Random Forest with this framework within this project’s mandate (Rigatti, 2017) (Ao, 2019).

### 4.1.4 Logistic Regression

LaValley (2008) made logistic regression, noted for its simplicity and interpretability, to serve as the benchmark model in this project. This method is indeed best suited for two class problems, and was used as a benchmark to assess if a linear model could perform as well in this dataset. Among them, the logistic regression model is of great importance in the medical science; the decision stated in the form must be comprehensible due to the fact that ‘the opacity is never tolerable’. This made it possible to gain simple and easily understandable ideas of the correlation between features and results and also pin point conditions that require advanced models. They were also checked to see how effective or otherwise they were with the diabetic retinopathy dataset. In case of analyzing individual models the performance is further enhanced by using the ensemble method (LaValley, 2008).

Where *Ŷ* is the estimated continuous outcome, is the linear regression equation for the independent variables in the model (Field, 2009).

To enhance the accuracy of the enriched features in the classification models, I used a soft voting classifier method (Gandhi, 2015). This classifier used the output from the SVM, Naïve Bayes, Random Forest, and Logistic Regression models. In this way, the Voting Classifier utilized the synergy effect when averaging the probability predictions of each model. For example, SVM is good for cases with clear margins, Random Forest for cases where feature interactions are significant, and Naïve Bayes for a quick reference. Developing such models together allowed the data to encompass a wider range of patterns that would improve prediction and reduce possible errors. This was especially advantageous since it reduced the influence that the validity and reliability of one model gave it over the presence or lack of predatory animals. This combination made it possible to introduce the features of each algorithm and to provide a stronger and more reliable work, which is especially crucial for medical applications such as DR diagnosis.

## 4.2 Model Evaluation Metrics

In order to measure the performance of each model as sharply and as inclusively as possible, several evaluation metrics were used. These metrics were selected because they give information about specific aspects of the models particularly due to class imbalance in the dataset. Accuracy was adopted as the first index, to assess the extent to which the models achieved correct classifications of samples. However, accuracy was insufficient in addressing imbalanced data since it was not designed to cater to the fragility of networks’ true positive or true negative identification.

Another was accuracy, this measured the percentage of cases that those that were predicted to have the positive outcome were actually right, with emphasis on the right positive cases. This metric is highly beneficial in medical environments due to the costs of a study misclassifying a patient without the condition as not having DR, for example, the risk of losing the opportunity to treat a patient. Likewise, recall or sensitivity was employed in establishing the discrete measure of how well the models identified true positive cases. This measure is important to heightened identification of all diabetic retinopathy patients who need to be treated because wrong negative results from the initial test will cost many patients their sight.

The traditional method of using precision and recall metrics complemented each other under one umbrella known as F1 score was used as it is useful when developing a model on imbalanced data sets. This metric was most useful during cases where both the false positives and the false negatives would be detrimental. Lastly, the ROC-AUC evaluated the ability of each model to classify data properly in terms of the thresholding angle. Since a higher AUC score ahs greater interpretability by signifying the models’ ability to classify with the right accuracy, positive and negative cases, the AUC was also a valuable overall test statistic. Such indicators (Yacouby, 2020) were essential for evaluating the models impartially and accurately. I also changed the evaluation function so that I kept the results as comparable as possible for different models with regards to interpretations that are relevant for deciding on the final choice of the model or changes to the model.

## 4.3 Hyperparameter Tuning

For better performance and refinement of the model, the hyperparameters were calibrated following the procedure explained by Schratz (2019). This process singled out on Random Forest since it has high baseline accuracy and is complex. Different configurations of parameters were tried out using the grid search method to arrived at the best configuration (Schratz, 2019).

Of all the changes made, number of trees was one of the parameters that was modified to make an improvement using the n\_estimators. A large number of trees increases accuracy by capturing many features about the data used in the construction of the model, though at the cost of time consumed. Another parameter, that is specific for Random Forest, is max\_depth, which determines the depth of every tree. While deeper trees allow analysis of more complicated relationships between the data components, it is easy to overtrain the trees and achieve a low level of accuracy on unseen data. Thus, the attempt was made to set this parameter in such a way that the model would achieve a good generalization while still having flexibility. The last parameter adjusted was min\_samples\_split that specifies the minimum number of samples required to make a split to the tree. By augmenting the value of this parameter, structure of the model’s trees are reduced such that overfitting is prevented and there is an improvement of its generality on new data. In combination, these changes allowed for developing a highly performing and stable Random Forest model.

In hyperparameter tuning the generalization capability of every configuration setting was ascertained through the three fold cross validation method that subdivides the data into three trains. This was beneficial in the first step to avoid the problem of overfitting since the performances of the model are tested on different divisions of the data to arrive at the final model. Hence, through fine-tuning of these parameters, I was able to attain the intended elevation of the generalization of the random forest model, as later cross-checked and compared with the walked-through as well as tuned version of the random forest model in terms of all four parameters.

# Chapter 5: Results

The results of this study are aimed at addressing the research questions by evaluating the performance of various machine learning models in predicting diabetic retinopathy (DR) using the Diabetic Retinopathy Debrecen dataset. The primary objective is to identify which model demonstrates the best performance across multiple metrics while ensuring that the analysis meets the project’s objectives.

From the evaluation of the Support Vector Machine (SVM) model, it can be concluded that the proposed model scored an overall accuracy of **71.55%**. From this, it could be seen that the designed model was able to classify the test cases with about **72%** accuracy. The accuracy for the SVM model was depending on the cases through it classified as positive for DR around **82%** were accurate. Recall on the other hand, which calculates the capability of the model to correctly identify all the actual positives was a little lower at **63.08%**. This means that the model failed to capture a lot of the positive cases as it was as low as **24**. However, the F-measure, the measure of precision and recall, was **71.30%**, which captured a fairly balanced measure of the model. The ROC-AUC score, which measures the potential of the model to separate between the positive and negative class, was **72.71%**. The Classification Report shows that there is high performance of the SVM in detecting negative samples (label 0) with high precision and recall values as compared to the positive samples (label 1). Although the experimental results are moderate, the SVM model proved its potential in the binary classification for DR but had some weaknesses in the recall that defines the system’s ability to identify all positive cases in a population.

The Naïve Bayes model yields an accuracy of 49.14%, anymore less than other models. It had 68.75 % accuracy in terms of probability, which means that, generally, it had a reasonable capacity to make correct estimates of positive cases it labeled. However, the recall was very poor at 16.92% meaning that the model hardly flagged all the true positives. It means that even F1 score was only 27.16%. Those included the overall performance measure of ROC-AUC of 53.56% which placed the models’ performance above just random chance. Working on the classification report we look that how the accuracy of the model is good for the negative cases but it is very poor for the positive cases. Possible reasons for this poor performance are stated below: As earlier discussed, the Naïve Bayes assumance of feature independence does not tally with this data set instance. Hence conversely although Naïve Bayes has an advantage of being computationally proficient they suffer from low accuracy and recall to be effectively used for DR prediction in this case.

For the predictor Random Forest, the accuracy rate was 72.41 % which proved to be better than required when compared to models like SVM and Naïve Bayes. In detail, the precision level for Random Forest was at 76.19%, and the recall level was at 73.85 %. These results show a desirable ratio of true positive cases detected and the rate of false positive results. This imbalance is captured fairly under the F1 score of 75 percent. The area under the receiver operating characteristic curve AUC-ROC was 72.22%, which confirm good discriminative ability of the constructed model. Based on the classification report, both positive and negative cases were well on predicting by the model with overall relatively balanced measures of precision, recall, and F-1 score measures of the two classes. Due to the nature of Random Forest as an ensemble of multiple decision trees, it was able to minimize overfitting while providing very good accuracy for this classification problem. Moreover, the feature importance analysis generated by the Random Forest model underscored critical predictive features, reinforcing its appropriateness for diabetic retinopathy prediction.

The analysis revealed that Logistic Regression outperformed SVM, Naïve Bayes, and Random Forest models with an accuracy of 75.86%. The model's precision was 83.64%, making the data highly reliable and the predictions positive. Its recall was 70.77%, indicating positive case detection. The F1 measure was 76.67%, indicating the model balanced precision and recall. The model's 76.56% ROC-AUC score was the highest of all models, indicating its greater effectiveness in class separation. The classification report shows that Logistic Regression performed well in both positive and negative classes, with the positive class performing slightly better. Logistic Regression is simple and clearer than other models for features and target variables, making it useful for medical applications.

Random Forest model and Logistic Regression model were tested with 70% and 75% accuracy respectively and the Hybrid Voting classifier used both the models, giving the best result. For this ensemble model the precision was 81.03 % and the recall was 72.31%. The obtained F1 score was 76.42% which proves that the proposed hybrid model is responsible for properly balancing precision and recall. Static using ROC-AUC score of 75.37% to validate how well it distinguished between positive cases and negative cases. The results of the classification of the RF and GB classifiers show that the Hybrid Voting Classifier achieved higher rates of precision and recall for both classes. The hybrid model was successful in combining the benefits and establishing high performance of the Random Forest and Logistic Regression algorithms. This proved more useful since Random Forest analysis’ strength in dealing with intricate interactions of features was complemented by Logistic Regression, which is more interpretable and stable. This performance profile of the hybrid model shows that the ensemble methods used in the current study have the possibility of providing better accuracy and model generalisation.

Table Result of Machine Learning Model

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ****Model**** | ****Accuracy**** | ****Precision**** | ****Recall**** | ****F1 Score**** | ****ROC-AUC**** |
| Support Vector Machine (SVM) | 71.55% | 82% | 63.07% | 71.3% | 72.71% |
| Naïve Bayes | 49.14% | 68.75% | 16.92% | 27.16% | 53.56% |
| Random Forest | 72.41% | 76.19% | **73.85%** | 75.00% | 72.22% |
| Logistic Regression | **75.86%** | **83.64%** | 70.77% | **76.67%** | **76.56%** |
| Hybrid Voting Classifier | 75.00% | 81.03% | 72.31% | 76.42% | 75.37% |

Among the evaluation metrics, accuracy provides an overview of the model's efficacy; however, it fails to account for class imbalance. Accuracy, recall, and precision are optimal metrics in this configuration, as these models are designed to identify true positives exclusively and to evade false positive occurrences. The F1 score, which balances precision and recall, is the most appropriate metric for addressing imbalanced data. ROC-AUC is favored for model fine-tuning as it encompasses all dimensions of the model's performance, making it suitable for model comparison.

This study unequivocally confirms that the selected models, particularly Logistic Regression and Hybrid Voting Classifier, exhibited the highest accuracy along with the maximum F1 and ROC-AUC scores. This paper addresses the research questions by demonstrating the efficacy of machine learning models in detecting diabetic retinopathy through features extracted from retinal images. Moreover, the application of ensemble methods enhances predictive accuracy, demonstrating their practical relevance. The insights gained from this project can facilitate the advancement of automated diabetic retinopathy detection systems, thereby enhancing early diagnosis and treatment for patients.

# Chapter 6: Analysis and Discussion

This study's findings clarify the issues concerning the application of machine learning models for assessing diabetic retinopathy (DR) using the Diabetic Retinopathy Debrecen dataset. This section examines the implications of these findings for the performance of each model, their relevance to the project's objectives, and their broader significance within the existing literature.

Logistic Regression emerged as the most efficient singular model in this analysis, achieving an accuracy of 75.86%. It attained the highest F1 score of 76.67% and a substantial ROC-AUC of 76.56%. These metrics indicate that Logistic Regression effectively conducted discriminant analysis between DR-positive and DR-negative cases, achieving a commendable balance of precision and recall. The efficacy of Logistic Regression can be attributed to its relative simplicity in utilizing coefficients to calculate outcome probabilities, as well as its efficiency in modeling binary classifiers where the feature variables are likely to exhibit a linear relationship with the output variable. It provided an investigable result that demonstrated the correlation between the identified features and the progression of DR. This aligns with our prior discussion from Emon (2021), which noted that Logistic Regression produced favorable outcomes, particularly with well-structured datasets such as the Diabetic Retinopathy Debrecen dataset. Nevertheless, the accuracy achieved in this study surpassed the reported in the literature (75%), indicating that meticulous preprocessing, feature selection, and hyperparameter optimization can enhance performance.

The Hybrid Voting Classifier, integrating Logistic Regression and Random Forest, displayed commendable performance with an accuracy of 75% and an F1 score of 76.42%. The ensemble model leveraged the strengths of both models; Logistic Regression is straightforward to interpret and explain heuristically, while Random Forest excels at capturing interactions among features. The Hybrid Voting Classifier demonstrated that ensemble models can be significantly more effective as their individual weaknesses mitigate those of other models during the voting process. The results of the hybrid model align with prior research, such as Odeh (2021), which indicated that ensemble methods possess the capability to improve the predictive outcomes of DR. The diminished accuracy may be ascribed to the complexities involved in model integration, which also serves to limit the models' generalizability.

The Random Forest model achieved an accuracy of 72.41% and an F1 score of 75%. Due to its efficacy in identifying significant features and reduced propensity for overfitting, it was a suitable candidate for diabetic retinopathy prediction. Employing the Random Forest modeling technique, we determined feature significance, specifically the exudate characteristics and optic disc size, to enhance the prediction of diabetic retinopathy (DR). This aligns with prior research, including Gadekallu (2020), which demonstrates that Random Forest surpasses the majority of machine learning algorithms in structured datasets. However, the marginally reduced accuracy compared to Logistic Regression may be attributed to the dataset's size and the features present, which might not optimally leverage the inherent capabilities of the Random Forest ensemble model.

The Support Vector Machine (SVM) model demonstrated a commendable accuracy of 71.55%, yet exhibited a relatively low recall rate of 63%. A significant proportion of patients with DR-positive tests were undetected by the model. Support Vector Machines (SVM) exhibit certain limitations, including the significant impact of kernel function selection on performance and sensitivity to feature scaling. Although it is efficient for binary classification tasks, its performance in this study casts doubt on the capacity to linearly segregate the selected feature space. This observation aligns with the literature, which indicates that SVM can exhibit high accuracy and perform effectively on specific datasets, but needs parameter tuning for optimal performance. The relative performance in this context is moderate, indicating that improved or more specific methodologies may be necessary to enhance its efficacy.

The Naïve Bayes model achieved a minimum accuracy of 49.14% and a minimum F1 score of 27.16%. Its low recall of 16.92%, indicating its ineffectiveness in yielding DR-positive results from the overall outcomes. The subpar performance can be attributed to the model's inadequacy in accounting for interactions between features, which is not applicable to this dataset. This eliminates Naïve Bayes for this task, as features associated with diabetic retinopathy, such as exudates and optic disc measurements, exhibit high correlation. While Naïve Bayes exhibits rapid computational speed and demonstrates efficacy with independent data, the findings of this study reveal its limitations in managing data with dependent features. This aligns with Emon (2021), which indicated that Naïve Bayes was outperformed by other models in predicting DR.

The results of this study correspond with the current literature in multiple aspects. Gadekallu (2020) and Gundluru (2022) analyzed and demonstrated that the implementation of deep learning, particularly the CNN model, has the capacity to enhance prediction accuracy and sensitivity of DR. These models, however, are complex and require substantial data and computational resources that are somewhat beyond the scope of this study. This project demonstrates that although traditional machine learning models do not achieve the performance levels of deep learning models, they can still perform adequately if data preprocessing is executed correctly, appropriate features are selected, and optimal hyperparameters are utilized. This is particularly crucial when resources are limited, necessitating the use of more fundamental models.

Consequently, the results of this study delineate certain specific limitations. While the Logistic Regression model and the Hybrid Voting Classifier established a solid benchmark, they rely on structured modeling and feature engineering techniques, which restrict their practical application in compounding and unstructured datasets. The dataset utilized in this study is pertinent and adequate for interpreting DR progression; however, it may not encompass additional details that could be elucidated through alternative methods such as optical coherence tomography (OCT). Additionally, a contributing factor that may have influenced the forecasting models is a limited and relatively undiversified data set, which poses risks in developing generalizable and precise models that could be informed by larger data sets.

Despite these limitations, the results align with the project's objectives and the established research questions. In addressing the initial research question, the study successfully identified models to predict diabetic retinopathy utilizing features derived from retinal images. This study also seeks to compare the performances of SVM, Naïve Bayes, Random Forest, Logistic Regression, and the Hybrid Voting Classifier to address the second research question. The concept of the hybrid model illustrates how the amalgamation of diverse models can improve predictive accuracy, addressing the final inquiry of the research.

In terms of practicality, the most effective models are Logistic Regression and Hybrid Voting Classifier. Owing to their high accuracy, F1 score, interpretability, and robustness, these algorithms are suitable for implementation in real-world applications. These models could be integrated into automated diabetic retinopathy screening programs to assist ophthalmologists in identifying patients with diabetic retinopathy for early intervention. This system could be particularly beneficial in regions lacking access to advanced imaging technology or specialist consultations.

In summary, the findings presented in this paper demonstrate the significant potential of machine learning models for diabetic retinopathy prediction. This study enhances the literature on automated disaster recovery detection and suggests the potential for improving prediction accuracy through a combination of various methodological approaches. Nonetheless, the results are promising, and future research should focus on incorporating larger sample sizes, employing deep-learning methodologies, and integrating additional optical images to enhance the accuracy and reliability of diabetic retinopathy prediction systems. Consequently, the results of this study may be utilized to develop recommendations that enhance patients' quality of life and alleviate the burden of diabetic retinopathy in healthcare environments.

# Conclusion

This study sought to employ machine learning models for diabetic retinopathy analysis utilizing the Diabetic Retinopathy Debrecen dataset. The two foremost models The two most effective models were Logistic Regression and the Hybrid Voting Classifier, achieving accuracies of 75.86% and 75.00%, respectively. These models yielded elevated F1 scores and ROC-AUC, indicating their proficiency in accurate classification and the selection of relevant feature subgroups for DR-positive and DR-negative datasets. The performance of Random Forests was marginally inferior; however, it remained quite dependable in identifying feature interactions and ranking feature importance for DR. Conversely, Naïve Bayes and Support Vector Machine (SVM) models exhibit comparatively lower accuracy, and recall is nearly unattainable, rendering them unsuitable for this dataset and data type.

The results of this study underline the potential of machine learning models in automating diabetic retinopathy identification. Logistic Regression was identified as a stable and easily interpretable model, whereas the Hybrid Voting Classifier demonstrated the benefits of employing multiple models. The results indicate that employing a conventional machine learning methodology, in conjunction with preprocessing, feature selection, and hyperparameter optimization, can yield competitive results. It is most advantageous in situations where advanced deep learning techniques cannot be applied, either due to insufficient data or resources.

The practical value of these models lies in their subsequent application within automated diabetic retinopathy screening systems. Such systems may assist ophthalmologists in conducting initial diagnostic assessments for diabetic retinopathy, enabling the identification of patients requiring further examination or treatment. It may be most beneficial in regions where expertise in these disease processes or specialized imaging is deficient. These models may alleviate the burden associated with vision loss from diabetic retinopathy by facilitating detection and prompt intervention.

While the study successfully identified and achieved particular objectives, there are distinct recommendations for future research. Initially, the dataset range and feature range could be expanded, thereby increasing the patient sample and potentially incorporating additional imaging modalities, such as optical coherence tomography (OCT). It was determined that larger and more extensive datasets would improve model generalization and stability. Secondly, future research may incorporate alternative models such as Convolutional Neural Networks (CNN), which have demonstrated superior performance compared to other models, albeit at a higher computational cost. Finally, the integration of supplementary temporal data to reflect the dynamic progression of DR could enhance the model for both identification and prediction.

In conclusion, the results of this analysis provide compelling evidence that machine learning techniques, specifically Logistic Regression and the Hybrid Voting Classifier, can effectively estimate diabetic retinopathy. The utilization of these models could significantly improve early diagnosis and aid caregivers in managing this disease. The proposed methods could be advanced and expanded for broader practical application through additional research and development to improve patient outcomes and overall health-related burdens.

# Bibliography

Gadekallu, T. R. (2020). Early detection of diabetic retinopathy using PCA-firefly based deep learning model. . *Electronics, 9(2),* , 274.

Emon, M. U. (2021). Performance analysis of diabetic retinopathy prediction using machine learning models. *In 2021 6th International Conference on Inventive Computation Technologies (ICICT). IEEE.*, pp. 1048-1052.

Odeh, I. M. (2021). Diabetic retinopathy detection using ensemble machine learning. *2021 international conference on information technology (ICIT). IEEE,* .

Gundluru, N. R. (2022). "Enhancement of detection of diabetic retinopathy using Harris hawks optimization with deep learning model. *Computational Intelligence and Neuroscience 2022.1* , 8512469.

Gadekallu, T. R. (2023). Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing* , 1-14.

Fong, D. S. (2003). Diabetic retinopathy. . *Diabetes care, 26*, s99-s102.

Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science* .

Yang, L. a. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing 415* , 295-316.

Antal, B. &. (2014). Diabetic Retinopathy Debrecen [Dataset]. *UCI Machine Learning Repository. https://doi.org/10.24432/C5XP4P.*

Camizuli, E. a. (2018). Exploratory data analysis (EDA). *The encyclopedia of archaeological sciences* , 1-7.

Tiessen, A. E.-R. (2017). Improved representation of biological information by using correlation as distance function for heatmap cluster analysis. *American Journal of Plant Sciences 8.3* , 502-516.

Menze, B. H. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics 10* , 1-16.

Çetin, V. a. (2022). A comprehensive review on data preprocessing techniques in data analysis. *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 28.2*, 299-312.

Vishwanathan, S. V. (2002). SSVM: a simple SVM algorithm. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 IEEE,* , Vol. 3.

Webb, G. I. (2010). Naïve Bayes. *Encyclopedia of machine learning 15.1* , 713-714.

Rigatti, S. J. (2017). Random forest. *Journal of Insurance Medicine 47.1 (2017):* , 31-39.

LaValley, M. P. (2008). Logistic regression. *Circulation 117.18* , 2395-2399.

Gandhi, I. a. (2015). Hybrid ensemble of classifiers using voting. *2015 international conference on green computing and Internet of Things (ICGCIoT). IEEE,* .

Yacouby, R. a. (2020). Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. *Proceedings of the first workshop on evaluation and comparison of NLP systems.* .

Schratz, P. e. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling 406* , 109-120.

Cortes, C. &. (1995). Support-vector networks. *Machine Learning, 20(3), doi:10.1007/bf00994018*, pp. 273–297.

McCallum, A. a. (1998). A comparison of event models for naive bayes text classification. *AAAI-98 workshop on learning for text categorization. Vol. 752. No. 1.* .

Ao, Y. H. (2019). The linear random forest algorithm and its advantages in machine learning assisted logging regression modeling. *Journal of Petroleum Science and Engineering 174* , 776-789.

Field, A. (2009). Logistic regression. *Discovering statistics using SPSS 264.1* , 315.

# Appendix

import pandas as diabetic\_pandas

from sklearn.preprocessing import StandardScaler as DiabeticScaler

from sklearn.model\_selection import train\_test\_split as Diabetic\_split, GridSearchCV

import matplotlib.pyplot as diabetic\_graph

from sklearn.ensemble import RandomForestClassifier, VotingClassifier

from sklearn.svm import SVC

from sklearn.metrics import accuracy\_score as acc\_diabetic, precision\_score as pre\_diabetic, recall\_score as re\_diabetic, f1\_score as f1\_diabetic, roc\_auc\_score, classification\_report as report\_diabetic

from sklearn.naive\_bayes import GaussianNB

from sklearn.linear\_model import LogisticRegression

from sklearn.model\_selection import GridSearchCV

file\_path = '/content/drive/My Drive/diabetic\_retinopathy.csv'

diabetic\_df = diabetic\_pandas.read\_csv(file\_path)

print("Dataset Preview:")

print(diabetic\_df.head())

print(diabetic\_df.describe())

diabetic\_df.head(7)

# Data Preprocessing: Handle missing values and normalize continuous features

print(diabetic\_df.isnull().sum())

# Normalizing

diabetic\_features = [

'exudate8', 'exudate7', 'exudate6',

'exudate3', 'exudate2', 'exudate1', 'exudate5',

'macula\_opticdisc\_distance', 'opticdisc\_diameter'

]

diabetic\_s = DiabeticScaler()

diabetic\_df [diabetic\_features] = diabetic\_s.fit\_transform(diabetic\_df [diabetic\_features])

Diabetic\_F = diabetic\_df.drop(columns=['Class'])

Diabetic\_L = diabetic\_df ['Class']

Diabetic\_F\_train, Diabetic\_F\_test, Diabetic\_L\_train, Diabetic\_L\_test = Diabetic\_split(Diabetic\_F, Diabetic\_L, test\_size=0.1, random\_state=42)

# Exploratory Data Analysis (EDA)

diabetic\_graph.figure(figsize=(10, 6))

sns.countplot(x='Class', data= diabetic\_df, palette='Set2')

diabetic\_graph.title("Class Distribution in Diabetic Retinopathy Dataset")

diabetic\_graph.show()

# Heatmap of correlations

diabetic\_graph.figure(figsize=(14, 10))

sns.heatmap(diabetic\_df.corr(), annot=False, cmap='coolwarm', linewidths=0.5)

diabetic\_graph.title("Correlation Heatmap")

diabetic\_graph.show()

# Feature Importance with Random Forest

# Feature Importance

rf\_diabetic\_model = RandomForestClassifier()

rf\_diabetic\_model.fit(Diabetic\_F\_train, Diabetic\_L\_train)

# Plot feature importance

feature\_imp = rf\_diabetic\_model.feature\_importances\_

imp\_df = diabetic\_pandas.DataFrame({'Feature': Diabetic\_F.columns, 'Importance': feature\_imp})

imp\_df = imp\_df.sort\_values(by='Importance', ascending=False)

diabetic\_graph.figure(figsize=(12, 8))

sns.barplot(x='Importance', y='Feature', data=imp\_df)

diabetic\_graph.title('Feature Importance from Random Forest')

diabetic\_graph.show()

# Select top N features based on importance

top\_features = imp\_df[imp\_df['Importance'] > 0.02]['Feature'].tolist() # Customize threshold

Diabetic\_F\_train\_selected = Diabetic\_F\_train[top\_features]

Diabetic\_L\_test\_selected = Diabetic\_F\_test[top\_features]

# Model Evaluation Function

def evaluate\_model(name, Diabetic\_L\_test, y\_pred):

print(f"Evaluation Metrics for {name}:")

print("Accuracy:", acc\_diabetic(Diabetic\_L\_test, y\_pred))

print("Precision:", pre\_diabetic(Diabetic\_L\_test, y\_pred))

print("Recall:", re\_diabetic(Diabetic\_L\_test, y\_pred))

print("F1 Score:", f1\_diabetic(Diabetic\_L\_test, y\_pred))

print("\nClassification Report:\n", report\_diabetic(Diabetic\_L\_test, y\_pred))

print("---------------------------------------------------")

# Kernels to test

kernels = ['linear', 'poly', 'rbf', 'sigmoid']

best\_kernel = None

best\_score = 0

# Loop through kernels

for k in kernels:

print(f"Testing SVM with k: {k}")

svm\_diabetic\_model = SVC(kernel=k, probability=True, random\_state=42)

svm\_diabetic\_model.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

y\_pred\_svm = svm\_diabetic\_model.predict(Diabetic\_F\_test\_selected)

# Evaluate the model

accuracy = acc\_diabetic(Diabetic\_L\_test, y\_pred\_svm)

print(f"Accuracy with k {k}: {accuracy}")

# Update the best k based on accuracy

if accuracy > best\_score:

best\_score = accuracy

best\_k = k

evaluate\_model(f"SVM with {k} k", Diabetic\_L\_test, y\_pred\_svm)

print(f"\nBest Kernel: {best\_k} with Accuracy: {best\_score}")

nb\_model = GaussianNB()

nb\_model.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

y\_pred\_nb = nb\_model.predict(Diabetic\_F\_test\_selected)

evaluate\_model("Naïve Bayes", Diabetic\_L\_test, y\_pred\_nb)

rf\_diabetic \_model = RandomForestClassifier(random\_state=42)

rf\_diabetic \_model.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

y\_pred\_rf = rf\_diabetic \_model.predict(Diabetic\_F\_test\_selected)

evaluate\_model("Random Forest", Diabetic\_L\_test, y\_pred\_rf)

lr\_diabetic\_model = LogisticRegression(random\_state=42, max\_iter=1000)

lr\_diabetic\_model.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

y\_pred\_lr = lr\_diabetic\_model.predict(Diabetic\_F\_test\_selected)

evaluate\_model("Logistic Regression", Diabetic\_L\_test, y\_pred\_lr)

param\_grid = {

'n\_estimators': [300, 400, 500],

'max\_depth': [None, 20, 10, 30],

'min\_samples\_split': [5, 2, 10]

}

grid\_diabetic\_search\_rf = GridSearchCV(rf\_diabetic\_model, param\_grid, cv=3, scoring='accuracy')

grid\_diabetic\_search\_rf.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

best\_rf\_model = grid\_diabetic\_search\_rf.best\_estimator\_

y\_pred\_best\_rf = best\_rf\_model.predict(Diabetic\_L\_test\_selected)

evaluate\_model("Best Random Forest (Tuned)", Diabetic\_L\_test, y\_pred\_best\_rf)

# Hyperparameter Logistic Regression

param\_grid = {

'C': [0.1, 0.01, 10, 100, 1],

'penalty': ['l2', 'l1', 'none', 'elasticnet'], # Regularization types

'max\_iter': [500, 1000, 100]

'solver': ['saga', 'liblinear'], # Solvers for optimization

}

grid\_diabetic\_search\_lr = GridSearchCV(LogisticRegression(random\_state=42), param\_grid, cv=3, scoring='accuracy')

grid\_diabetic\_search\_lr.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

# Logistic Regression

best\_lr\_model = grid\_diabetic\_search\_lr.best\_estimator\_

y\_pred\_best\_lr = best\_lr\_model.predict(Diabetic\_F\_test\_selected)

evaluate\_model("Best Logistic Regression (Tuned)", Diabetic\_L\_test, y\_pred\_best\_lr)

# Hybrid Model: Voting Classifier (Random Forest, and Logistic Regression)

hybrid\_model = VotingClassifier(

estimators=[

('rf', best\_rf\_model),

('lr', lr\_diabetic\_model)

],

voting='soft'

)

hybrid\_model.fit(Diabetic\_F\_train\_selected, Diabetic\_L\_train)

y\_pred\_hybrid = hybrid\_model.predict(Diabetic\_F\_test\_selected)

evaluate\_model("Hybrid Voting Classifier", Diabetic\_L\_test, y\_pred\_hybrid)