Ocular Disease Detection Using Machine Learning

Project/Thesis ID. 2023: 111

Session: BSc. Spring 2001

Project Supervisor: Dr. Sarmad Shams

Submitted By

Mishkaat Jamil(20BME023)

Aqsa Faheem(20BME009)

Afnan Qureshi(20BME002)

Zona Khan(20BME038)

Institute of Biomedical Engineering & Technology

Liaquat University of Medical & Health Sciences, Jamshoro

Certification

This is to certify that **Mishkaat Jamil(20BME023), Aqsa Faheem(20BME009), Afnan Qureshi(20BME002) & Zona Khan(20BME038)** have successfully completed the final project **Ocular Disease Detection Using Machine Learning,** at the **Liaquat University of Medical and Health Sciences Jamshoro**, to fulfill the partial requirement of the degree **Bs Biomedical Engineering and Technology.**

External Examiner	**Project Supervisor**
[Name of Examiner]	Dr. Sarmad Shams
[Designation]	Head of Department

Dr. Sarmad Shams

Chairman

Department of [Insitute of Biomedical Engineering & Technology], [LUMHS]

Project Title (Ocular Disease Detection Using Machine Learning)

Sustainable Development Goals

(Please tick the relevant SDG(s) linked with FYDP)

SDG No	Description of SDG	SDG No	Description of SDG
SDG 1	No Poverty	SDG 9	Industry, Innovation, and Infrastructure
SDG 2	Zero Hunger	SDG 10	Reduced Inequalities
SDG 3	Good Health and Well Being	SDG 11	Sustainable Cities and Communities
SDG 4	Quality Education	SDG 12	Responsible Consumption and Production
SDG 5	Gender Equality	SDG 13	Climate Change
SDG 6	Clean Water and Sanitation	SDG 14	Life Below Water
SDG 7	Affordable and Clean Energy	SDG 15	Life on Land
SDG 8	Decent Work and Economic Growth	SDG 16	Peace, Justice and Strong Institutions
		SDG 17	Partnerships for the Goals

Range of Complex Problem Solving			
	Attribute	**Complex Problem**	
1	Range of conflicting	Involve wide-ranging or conflicting technical, engineering and other issues.	

	requirements		
2	Depth of analysis required	Have no obvious solution and require abstract thinking, originality in analysis to formulate suitable models.	
3	Depth of knowledge required	Requires research-based knowledge much of which is at, or informed by, the forefront of the professional discipline and which allows a fundamentals-based, first principles analytical approach.	
4	Familiarity of issues	Involve infrequently encountered issues	
5	Extent of applicable codes	Are outside problems encompassed by standards and codes of practice for professional engineering.	
6	Extent of stakeholder involvement and level of conflicting requirements	Involve diverse groups of stakeholders with widely varying needs.	
7	Consequences	Have significant consequences in a range of contexts.	
8	Interdependence	Are high level problems including many component parts or sub-problems	
Range of Complex Problem Activities			
	Attribute	**Complex Activities**	
1	Range of resources	Involve the use of diverse resources (and for this purpose, resources include people, money, equipment, materials, information and technologies).	
2	Level of interaction	Require resolution of significant problems arising from interactions between wide ranging and conflicting technical, engineering or other issues.	
3	Innovation	Involve creative use of engineering principles and research-based knowledge in novel ways.	
4	Consequences to society and the environment	Have significant consequences in a range of contexts, characterized by difficulty of prediction and mitigation.	
5	Familiarity	Can extend beyond previous experiences by applying principles-based approaches.	

Abstract

Machine learning is crucial in helping medical professionals identify diseases early. Ophthalmic disorders are often not life-threatening, but when they advance over time, they may have a major influence on the patient's quality of life. Early detection is essential for avoiding serious consequences and for enhancing patient outcomes. The ultimate goal of this research is to develop a precise and trustworthy tool that may help healthcare professionals in the early identification and monitoring of ocular disorders. This system uses a dataset made up of diverse ocular photographs. The successful implementation of this system may improve patient outcomes by enabling quick interventions and reducing the risk of vision loss brought on by ocular disorders. In order to create a reliable system, this project integrates the fields of biomedical engineering and machine learning. Machine learning is revolutionizing ocular eye disease detection by leveraging datasets of eye images and algorithms. The process involves acquiring diverse eye image data, preprocessing it to enhance quality, and extracting relevant features such as blood vessel patterns and optic disc characteristics. Machine learning models are then trained using these features. The models are evaluated using separate datasets, assessing their accuracy, sensitivity, specificity, and overall performance. These models can be used in clinical settings once they have been validated to enhance the knowledge of eye care specialists

Keywords: Machine Learning; Ocular Disease; Algorithms; Accuracy; Opthalmic disorders.

Undertaking

We certify that the project **Ocular Disease Detection Using Machine Learning** is our own work. The work has not, in whole or in part, been presented elsewhere for assessment. Where material has been used from other sources it has been properly acknowledged/ referred.

Mishkaat Jamil

20BME023

Aqsa Faheem

20BME009

Afnan Qureshi

20BME002

Zona Khan

20BME38

Acknowledgement

We truly acknowledge the cooperation and help make by [**Dr. Sarmad Shams], Designation** of **[Head of Department].** He has been a constant source of guidance throughout the course of this project. We would also like to thank [**Engr. Natasha Mukhtiar**] from [**Lecturer]**, [**IBET** ,**LUMHS]** for his help and guidance throughout this project.

We are also thankful to our friends and families whose silent support led us to complete our project.

Table of Contents

Chapter 1

1.1 Introduction

Pakistan is the sixth most populous country in the world. Approximately 67% of total population lives in rural areas, according to World Health Organization estimates, there are 177 million diabetics in the world. One in 20 of world's adult population now suffers from diabetes. At least one in ten deaths among adults between 35 and 64 years old is related to diabetes. It is estimated that by the year 2030, diabetic population of the world will be doubled. People suffering from diabetes are at high risk of developing various ocular disease over time. Ocular diseases present a substantial global health challenge, impacting millions of individuals worldwide. Among these conditions, Diabetic Eye Disease (DED) is of particular concern, encompassing various disorders like Diabetic Retinopathy, Diabetic Macular Edema, Glaucoma, and Cataracts.

1.1.1 Diabetes Mellitus:

Diabetes mellitus, commonly referred to as diabetes, is a chronic metabolic disorder characterized by elevated levels of glucose (sugar) in the blood a condition called Hyperglycemia (diabetes paper). Diabetes is a disease that affects the body's ability to produce or use insulin effectively to control blood sugar (glucose) levels. Too much glucose in the blood for a long time can cause damage in many parts of the body. Diabetes can damage the heart, kidneys and blood vessels. It damages small blood vessels in the eye as well. Even if diabetes is well controlled, it can affect your regular eye care.

Numerous physiological organs are impacted by hyperglycemia and the related protein, lipid, and carbohydrate metabolic dysfunctions, which prevent them from operating normally. Organ damage, dysfunction, and, ultimately, organ failure characterizes these complications and affect body organs, which include, in particular, eyes, kidneys, heart, and nerves. Eye-related complications result in retinopathy with progression to blindness.

1.1.1.1Types Of Diabetes Mellitus:

There are two main types of Diabetes Mellitus.

1. **Type 1 Diabetes Mellitus:**

T1DM, also known as type 1A DM or as per the previous nomenclature as insulin-dependent diabetes mellitus (IDDM) constitutes about 5–10% of all the cases of diabetes. It is an autoimmune disorder characterized by T-cell-mediated destruction of pancreatic β-cells, which results in insulin deficiency and ultimately hyperglycemia. (diabetes fyp). Individuals with type 1 diabetes have little to no insulin production and require lifelong insulin therapy.

2. **Type 2 Diabetes Mellitus:**

T2DM, also known as non-insulin-dependent diabetes mellitus (NIDDM) or adult onset diabetes, as per the previous nomenclature, constitutes about 90–95% of all the cases of diabetes. They are two main insulin-related abnormalities that are present in this kind of diabetes: insulin resistance and dysfunctional beta cells. (diabetes fyp). Type 2 diabetes is often associated with lifestyle factors such as obesity, physical inactivity, and an unhealthy diet.

1.1.1.2 Causes of Diabetes:

- **Genetics:** Diabetes can run in families due to genetics. You can be genetically prone to the disease if diabetes runs in your family. Your risk of developing diabetes may increase due to specific genes.
- **Insulin Resistance:** People with type 2 diabetes have cells that don't react well to insulin. Insulin resistance is what causes this. It frequently goes hand in hand with obesity, inactivity, and a bad diet.
- **Autoimmune Reaction:** In type 1 diabetes, the immune system unintentionally targets and kills the beta cells in the pancreas that produce insulin. The autoimmune reaction causes the body to produce less insulin.
- **Hormonal Changes:** Gestational diabetes can appear in some pregnant women. Elevated blood sugar levels can result from hormonal changes during pregnancy that alter insulin's ability to function.

- **Obesity:** Being overweight or obese is a significant risk factor for type 2 diabetes. Excess body fat, especially around the abdomen, can contribute to insulin resistance.

3. Diabetic Eye Disease:

Diabetic eye disease is a term for several eye problems that can all result from diabetes. Diabetic eye disease includes:

- diabetic retinopathy,
- cataract, and
- glaucoma.

1. Diabetic Retinopathy:

A microvascular condition known as diabetic retinopathy (DR) can develop in people with diabetes mellitus. It is a chronic eye condition that damages the retina, the tissue at the back of the eye that is sensitive to light. DR is a substantial contributor to vision loss and blindness, especially in working-age individuals. Long-term diabetes destroys the blood vessels in the retina, causing the disease to develop. If this injury is not identified and treated right once, it might cause visual loss

1. Pathophysiology:

Long-term hyperglycemia (high blood sugar levels) destroys the small blood vessels that nourish the retina, which is the main cause of DR. Due to this damage, the blood vessels start to leak, which causes the retinal tissue to enlarge and impairs vision. In advanced stages, aberrant blood vessels may develop on the retina's surface, which might cause blindness or severe visual loss.

2. Diagnostic Techniques:

Various diagnostic techniques have been created over time to recognize and keep track of diabetic retinopathy. Ophthalmologists frequently employ fluorescein angiography, optical coherence tomography (OCT), and fundus photography to determine the degree of retinal damage. These techniques offer precise retinal pictures that enable medical experts to assess the severity of the problem

3. Epidemiology:

In line with the global growth in diabetes patients, diabetic retinopathy prevalence has been gradually rising. According to estimates, one-third of all patients with diabetes mellitus have diabetic retinopathy to some extent. The length of diabetes, glycemic control, and other systemic variables including hypertension

4. Treatment Modalities:

There are several treatment options for diabetic retinopathy, ranging from lifestyle changes and medication to surgical procedures. Diabetes identification and control, as well as frequent eye exams, are critical in preventing disease development. Some of the main therapies used to preserve or restore vision in afflicted patients include laser therapy, anti-VEGF (vascular endothelial growth factor) injections, and vitrectomy.

2. Cataract:

One of the most common age-related ocular disorders, cataracts, is defined by clouding the natural lens of the eye. This disorder has a serious influence on vision and is the primary cause of blindness worldwide. Regular eye care requires an understanding of the historical setting, etiology, diagnostic techniques, and available treatments for cataracts.

1. Pathophysiology:

Cataracts arise when proteins in the lens of the eye start to clump together, creating hazy patches that block the flow of light. Because of lens clouding, one may have blurry or distorted vision, trouble seeing in low light, and susceptibility to glare. Although other variables such as diabetes, exposure to UV radiation, and some drugs might hasten its development, the process is slow and mostly linked to age.

2. Diagnostic Techniques:

The identification of cataracts has been improved using modern diagnostic techniques. To determine the degree of lens opacity, ophthalmologists use methods including slit-lamp examination, visual acuity testing, and optical coherence tomography (OCT). By allowing for an accurate diagnosis, these techniques help medical personnel choose the best course of action for patients, whether that be the use of corrective lenses or surgical removal of the cataract.

3. Epidemiology:

Cataracts impact millions of individuals worldwide, affecting people of all ages. However, aging is most frequently associated with cataracts. Age-related increases in cataract prevalence make the condition a serious public health concern in nations with aging populations. The prevalence of cataracts is also influenced by socioeconomic variables, availability of healthcare, and genetic susceptibility.

Treatment Modalities:

The most successful form of treatment for cataracts is cataract surgery, which involves the removal of the clouded lens and its replacement with an artificial intraocular lens (IOL). Cataract surgery has evolved into one of the safest and most frequently performed operations in the world because of improvements in surgical methods and IOL technology

3. Glaucoma:

A series of eye diseases known as glaucoma harm the optic nerve, which, if unchecked, results in irreversible vision loss. It is one of the main contributors to blindness globally. Although the condition is frequently linked to increased intraocular pressure (IOP), normal-tension glaucoma can also develop with normal or low IOP. Typically, glaucoma proceeds slowly and painlessly, and the progressive loss of vision begins with peripheral vision.

There are several forms of glaucoma, which may be roughly divided into open-angle and angle-closure forms. The most prevalent kind of glaucoma is open-angle glaucoma, in which the drainage aperture of the eye is left open but the aqueous humor does not effectively exit the eye. On the other side, angle-closure glaucoma happens when the iris covers the drainage angle, causing a sharp rise in IOP.

1. **Pathophysiology:**

Elevated intraocular pressure is frequently associated with glaucoma, which is essentially defined by gradual damage to the optic nerve head. The precise mechanisms causing injury to the optic nerve are intricate and complicated. Although increased IOP is a key risk factor, glaucoma is also significantly influenced by other variables, including genetics, vascular factors, and mechanical issues with the optic nerve head.

2. **Diagnostic Techniques**:

Intraocular pressure measurement, an evaluation of the optic nerve head, an assessment of the visual field, and occasionally imaging procedures like optical coherence tomography (OCT) are all necessary for the diagnosis of glaucoma. Regular eye exams are crucial for early detection, especially for people who have risk factors including a family history of glaucoma.

3. **Epidemiology:**

Glaucoma is a widespread eye condition with significant implications for public health. It stands as one of the leading causes of irreversible blindness globally. As age increases, so does the prevalence of glaucoma, making it particularly common among older individuals. While gender does not significantly influence glaucoma's prevalence, certain racial and ethnic groups, such as African Americans and Hispanics, are at a higher risk of developing the condition. Elevated intraocular pressure (IOP) is a primary risk factor for glaucoma, though not all individuals with high IOP develop the condition

4. Treatment Modalities:

Although glaucoma cannot be cured, early detection and treatment can help manage the condition and stop future visual loss. In circumstances when drugs and laser therapy are inadequate, surgical treatments such as trabeculectomy or the implantation of drainage devices are possibilities for treatment.

1.2 Statement of the problem

Diabetic eye diseases pose a significant health risk to individuals with diabetes. Early detection is the key challenge in healthcare. Current screening techniques usually rely on subjective evaluations or poor diagnostic tools which delays diagnoses and constricts the range of available treatments. This delay might result in serious issues for individuals like irreversible loss of vision. The high cost and restricted accessibility of modern diagnostic equipment further hinder the early treatment of eye conditions.

1.3 Goals/Aims & Objectives

Ocular disease detection refers to the process of identifying and diagnosing diseases and conditions that affect the eye and its surrounding structures. The goals of ocular disease detection using machine learning involve enhancing early detection, improving accuracy and efficiency, personalizing treatment, enabling remote monitoring, and advancing research efforts, all of which can ultimately lead to better patient outcomes and a reduction in visual impairment caused by ocular diseases. The proposed system will help people to get the proper treatment of the mentioned diseases at an early stage thus reducing the percentage of blindness being caused. The results we are going to achieve through this project are the detection of diabetic eye diseases which includes Diabetic Retinopathy, Diabetic Macular Edema, Glaucoma & Cataract.

Chapter 2

2.1 ALGORITHMS

o k-Nearest Neighbor

(kNN) algorithm is an effortless but productive machine learning algorithm. It is effective for classification as well as regression. However, it is more widely used for classification prediction. kNN groups the data into coherent clusters or subsets and classifies the newly inputted data based on its similarity with previously trained data. The input is assigned to the class with which it shares the nearest neighbors. Though kNN is effective, it has many weaknesses. This paper highlights the kNN method and its modified versions available in previously done research. These variants remove the weaknesses of kNN and provide a more efficient method.

i.e. it does not make any presumptions on the elementary dataset. It is known for its simplicity and effectiveness. It is a supervised learning algorithm. A labeled training dataset is provided where the data points are categorized into various classes, so that the class of the unlabeled data can be predicted. In Classification, different characteristics determine the class to which the unlabeled data belongs. KNN is mostly used as a classifier. It is used to classify data based on closest or neighboring training examples in a given region. This method is used for its simplicity of execution and low computation time. For continuous data, it uses the Euclidean distance to calculate its nearest neighbors. For a new input, the K nearest neighbors are calculated and the majority among the neighboring data decides the classification for the new input. Even though this classifier is simple, the value of ‘K’ plays an important role in classifying the unlabeled data. There are many ways to decide the values for ‘K’, but we can simply run the classifier multiple times with different values to see which value gives the most effective result. The computation cost is slightly high because all the calculations are made when the training data is being classified, not when it is encountered in the dataset. It is a lazy learning algorithm as not much is done when the dataset is being trained except storing the training data and memorizing the dataset instead. It does not perform generalization on the training dataset. So the entire fundamental dataset being trained is required when in the testing

stage. In regression, KNN predicts continuous values. This value is the average of the values of its K - nearest neighbors. KNN is used in datasets where data is separated into different clusters so that the class of the new input can be determined. KNN is more significant for a study where there is no previous knowledge about the data being used. also, K-NN K-nearest-neighbor classification was developed to execute characteristic analysis when clear parametric approximations of probability densities were unknown or difficult to determine. In an unpublished US Air Force School of Aviation

Medicine report in 1951, Fix and Hodges introduced a non-parametric algorithm for pattern classification that has since become known the Knearest neighbor rule. A. **WORKING:** k-NN is a classification algorithm. Mainly there are two steps in classification:

1. Learning Step: Using the training data a classifier is constructed.

2. Assessment of the classifier.

According to the nearest neighbor technique, the new unlabeled data is classified by determining which classes its neighbors belong to. KNN algorithm utilizes this concept in its calculation. In the case of KNN algorithm, a particular value of K is fixed which helps us in classifying the unknown tuple. When a new unlabeled tuple is encountered in the dataset, KNN performs two operations. First, it analyzes the K points closest to the new data point, i.e. the K nearest neighbors. Second, using the neighbors' classes, KNN determines as to which class the new data should be classified into. A simple KNN When some new data is added, it classifies the data accordingly. It is more useful in a dataset which is roughly divided into clusters and belongs to a specific region of the data plot. Thus this algorithm brings more accuracy in dividing the data inputs into different classes in a clearer way. KNN figures out the class having the maximum number of points sharing the least distance from the data point that needs to be classified. Hence, the Euclidean distance needs to be calculated between the test sample and the specified training samples.

- **SVM**

  Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Support Vector Machine Terminology

1. **Hyperplane:** Hyperplane is the decision boundary that is used to separate the data points of different classes in a feature space. In the case of linear classifications, it will be a linear equation i.e. wx+b = 0.
2. **Support Vectors:** Support vectors are the closest data points to the hyperplane, which makes a critical role in deciding the hyperplane and margin.
3. **Margin**: Margin is the distance between the support vector and hyperplane. The main objective of the support vector machine algorithm is to maximize the margin. The wider margin indicates better classification performance.
4. **Kernel**: Kernel is the mathematical function, which is used in SVM to map the original input data points into high-dimensional feature spaces, so, that the hyperplane can be easily found out even if the data points are not linearly separable in the original input space. Some of the common kernel functions are linear, polynomial, radial basis function(RBF), and sigmoid.
5. **Hard Margin:** The maximum-margin hyperplane or the hard margin hyperplane is a hyperplane that properly separates the data points of different categories without any misclassifications.

6. **Soft Margin:** When the data is not perfectly separable or contains outliers, SVM permits a soft margin technique. Each data point has a slack variable introduced by the soft-margin SVM formulation, which softens the strict margin requirement and permits certain misclassifications or violations. It discovers a compromise between increasing the margin and reducing violations.
7. **C:** Margin maximisation and misclassification fines are balanced by the regularization parameter C in SVM. The penalty for going over the margin or misclassifying data items is decided by it. A stricter penalty is imposed with a greater value of C, which results in a smaller margin and perhaps fewer misclassifications.
8. **Hinge Loss:** A typical loss function in SVMs is hinge loss. It punishes incorrect classifications or margin violations. The objective function in SVM is frequently formed by combining it with the regularisation term.
9. **Dual Problem:** A dual Problem of the optimisation problem that requires locating the Lagrange multipliers related to the support vectors can be used to solve SVM. The dual formulation enables the use of kernel tricks and more effective computing.

I. Types of Support Vector Machine

Based on the nature of the decision boundary, Support Vector Machines (SVM) can be divided into two main parts:

- **Linear SVM:** Linear SVMs use a linear decision boundary to separate the data points of different classes. When the data can be precisely linearly separated, linear SVMs are very suitable. This means that a single straight line (in 2D) or a hyperplane (in higher dimensions) can entirely divide the data points into their respective classes. A hyperplane that maximizes the margin between the classes is the decision boundary.
- **Non-Linear SVM:** Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher-dimensional feature space, where the data points can be linearly separated. A

linear SVM is used to locate a nonlinear decision boundary in this modified space.

○ DECISION TREE

A decision tree is a flowchart-like tree structure where each internal node denotes the feature, branches denote the rules and the leaf nodes denote the result of the algorithm. It is a versatile supervised machine-learning algorithm, which is used for both classification and regression problems. It is one of the very powerful algorithms. It is also used in Random Forest to train on different subsets of training data, which makes Random Forest one of the most powerful algorithms in machine learning

- **Root Node:** It is the topmost node in the tree, which represents the complete dataset. It is the starting point of the decision-making process.
- Decision/Internal Node: A node that symbolizes a choice regarding an input feature. Branching off of internal nodes connects them to leaf nodes or other internal nodes.
- **Leaf/Terminal Node:** A node without any child nodes that indicates a class label or a numerical value.
- **Splitting:** The process of splitting a node into two or more sub-nodes using a split criterion and a selected feature.
- **Branch/Sub-Tree:** A subsection of the decision tree starts at an internal node and ends at the leaf nodes.
- **Parent Node:** The node that divides into one or more child nodes.
- **Child Node:** The nodes that emerge when a parent node is split.
- **Impurity**: A measurement of the target variable's homogeneity in a subset of data. It refers to the degree of randomness or uncertainty in a set of examples. The **Gini index** and **entropy** are two commonly used impurity measurements in decision trees for classification task
- **Variance**: Variance measures how much the predicted and the target variables vary in different samples of a dataset. It is used for regression problems in decision trees. **Mean squared error, Mean Absolute Error, friedman_mse, or**

Half Poisson deviance are used to measure the variance for the regression tasks in the decision tree.

- **Information Gain:** Information gain is a measure of the reduction in impurity achieved by splitting a dataset on a particular feature in a decision tree. The splitting criterion is determined by the feature that offers the greatest information gain, It is used to determine the most informative feature to split on at each node of the tree, with the goal of creating pure subsets
- **Pruning**: The process of removing branches from the tree that do not provide any additional information or lead to overfitting.

How does the Decision Tree algorithm Work?

The decision tree operates by analyzing the data set to predict its classification. It commences from the tree's root node, where the algorithm views the value of the root attribute compared to the attribute of the record in the actual data set. Based on the comparison, it proceeds to follow the branch and move to the next node.

The algorithm repeats this action for every subsequent node by comparing its attribute values with those of the sub-nodes and continuing the process further. It repeats until it reaches the leaf node of the tree. The complete mechanism can be better explained through the algorithm given below.

- Step-1: Begin the tree with the root node, says S, which contains the complete dataset.
- Step-2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Step-3: Divide the S into subsets that contains possible values for the best attributes.
- Step-4: Generate the decision tree node, which contains the best attribute.
- Step-5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node Classification and Regression Tree algorithm.

Advantages of the Decision Tree:

1. It is simple to understand as it follows the same process which a human follows while making any decision in real-life.

2. It can be very useful for solving decision-related problems.
3. It helps to think about all the possible outcomes for a problem.
4. There is less requirement of data cleaning compared to other algorithms.

o CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN), also called ConvNet, is a type of Artificial Neural Network(ANN), which has deep feed-forward architecture and has amazing generalizing ability as compared to other networks with FC layers, it can learn highly abstracted features of objects especially spatial data and can identify them more efficiently. A deep CNN model consists of a finite set of processing layers that can learn various features of input data (e.g. image) with multiple level of abstraction . The initiatory layers learn and extract the high level features (with lower abstraction), and the deeper layers learns and extracts the low level features (with higher abstraction). The basic conceptual model of CNN was shown in figure 2, different types of layers described in subsequent sections.

• One of the main reason for considering CNN in such case is the weight sharing feature of CNN, that reduce the number of trainable parameters in the network, which helped the model to avoid overfitting and as well as to improved generalization. • In CNN, the classification layer and the feature extraction layers learn together, that makes the output of the model more organized and makes the output more dependent to the extracted features.

• The implementation of a large network is more difficult by using other types of neural networks rather than using Convolutional Neural Networks. Nowadays CNN has been emerged as a mechanism for achieving promising result in various computer vision based applications like image classification, object detection, face detection, speech recognition, vehicle recognition, facial expression recognition, text recognition and many more

we have described the basic concepts of convolutional neural network (CNN) as well as the different key components of CNN architecture. Here in this section we try to discuss the training or learning process of a CNN model with certain guidelines in order to reduce the required training time and to improve model accuracy. The

training process mainly includes the following steps: • Data pre-processing and Data augmentation. • Parameter initialization. • Regularization of CNN. • Optimizer selection.

○ RANDOM FOREST

Random forest algorithms have three main hyper parameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote—i.e. the most frequent categorical variable—will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.
Key benefits for using this algorithm in our project

- Reduced risk of overfitting: Decision trees run the risk of overfitting as they tend to tightly fit all the samples within training data. However, when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error.
- Provides flexibility: Since random forest can handle both regression and classification tasks with a high degree of accuracy, it is a popular method among data scientists. Feature bagging also makes the random forest classifier

an effective tool for estimating missing values as it maintains accuracy when a portion of the data is missing.

- Easy to determine feature importance: Random forest makes it easy to evaluate variable importance, or contribution, to the model. There are a few ways to evaluate feature importance. Gini importance and mean decrease in impurity (MDI) are usually used to measure how much the model's accuracy decreases when a given variable is excluded. However, permutation importance, also known as mean decrease accuracy (MDA), is another important measure. MDA identifies the average decrease in accuracy by randomly permitting the feature values in oob samples.

Chapter 4

4.1 Proposed Solution/Results & Discussion

Accuracy:

The accuracy of algorithms used i-e Support vector Machine, Random Forest, Deecision Tree & KNN are shown in Table 1.

Algorithms	Accuracy
Support Vector Machine	80%
Random Forest	77%
Decision Tree	75%
KNN	71%

Table 1: Algorithms and their Accuracy

Predictions:

- The results predicted by Support vector machine are shown below:

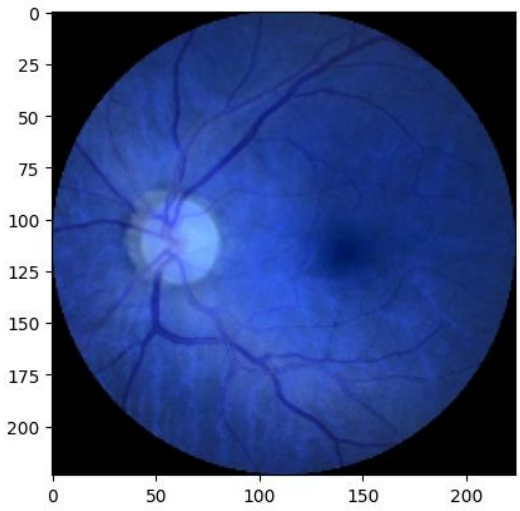

Figure 1 showing cataract predicted by SVM

- The results predicted by Random forest are shown below:

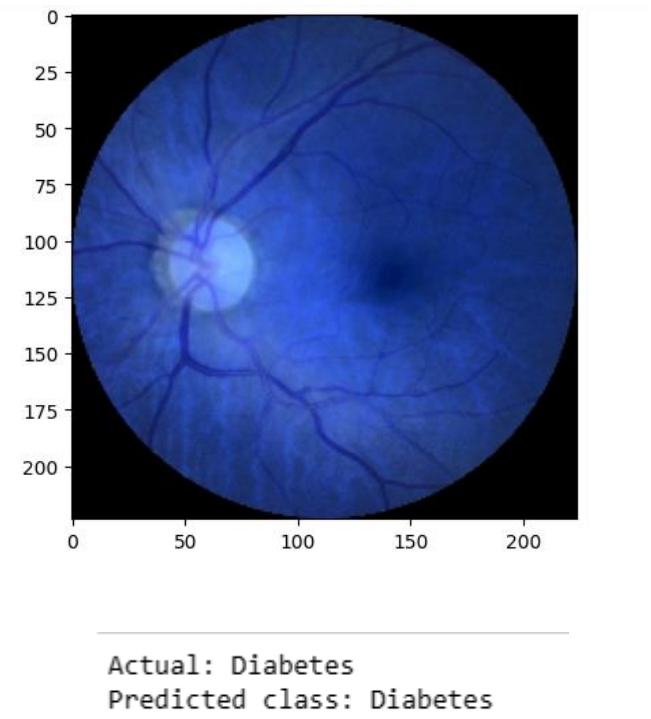

Figure 2 showing Diabetes predicting by Random forest

- The results predicted by Decision tree are shown below:

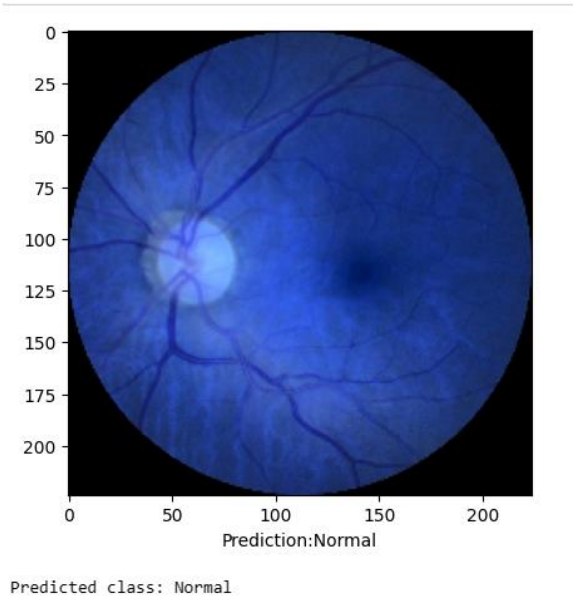

Figure 3showing Normal predicting by Decision tree

- The results predicted by KNN are shown below:

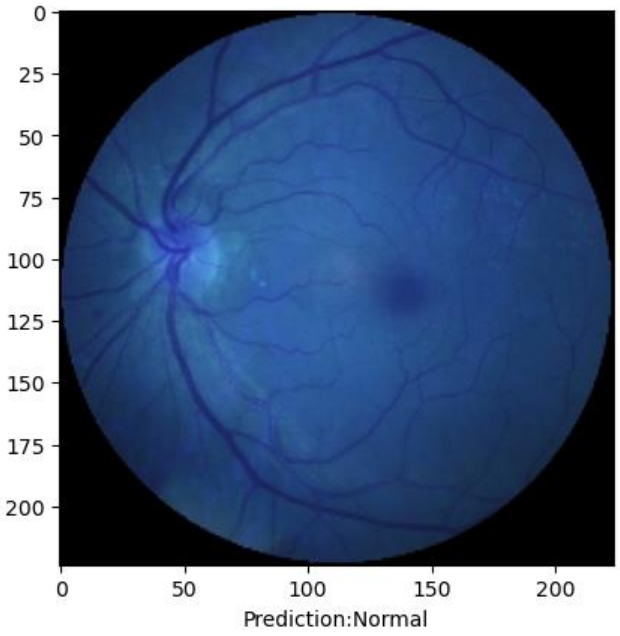

Figure 4showing Normal predicting by KNN

Chapter 5

5.1 Summary and Future work

In our final year project on ocular disease detection, we meticulously designed a robust system harnessing the power of machine learning. The project revolves around five key algorithms—KNN, SVM, decision tree, random forest, and CNN—each contributing unique insights to enhance the accuracy of disease identification.

We meticulously curated and preprocessed diverse ocular datasets to train and evaluate these algorithms comprehensively. The utilization of KNN and SVM underscores the significance of pattern recognition and classification, while decision trees and random forests bring a nuanced understanding of complex data relationships. The integration of CNN, with its deep learning capabilities, further refines the detection process by capturing intricate features within ocular images.

Our project doesn't merely stop at algorithm implementation; we conducted rigorous testing and validation to ensure the reliability of the detection system. The fusion of multiple algorithms provides a more holistic approach, minimizing false positives and false negatives. The results obtained showcase the potential of machine learning in revolutionizing ocular disease diagnostics.

Looking forward, future work could involve fine-tuning the existing algorithms, exploring ensemble methods to leverage their collective strength, and incorporating additional datasets to enhance the model's generalizability. The practical application of the system in real-world scenarios, alongside considerations for user interface development, marks the next steps towards translating our research into tangible impact within the medical field.