Diabetic Detection Using Irish

Dr G.Revathy,
Assistant Professor,
Erode Sengunthar Engineering College(Autonomous),
Perundurai, Erode.

Ms Anju, Parimalam and Imaya Kanishka,
Final CSE,
Erode Sengunthar Engineering College(Autonomous),
Perundurai, Erode.

Abstract

DiabeticIrish(DI)ishumaneyediseaseamongpeoplewithdiabeticswhichcauses damage to irish of eye and may eventually lead to complete blindness. Detection of diabetic Irishinearlystageisessentialtoavoidcompleteblindness.EffectivetreatmentsforDIare available though it requires early diagnosis and the continuous monitoring of diabetic patients. Also many physical tests like visual acuity test, pupil dilation, and optical coherence tomography canbeusedtodetectdiabeticIrishbutaretimeconsuming.Theobjectiveofourthesisisto givedecisionaboutthepresenceofdiabeticIrishbyapplyingensembleofmachinelearning classifying algorithms on features extracted from output of different irishl image. It will give us accuracy of which algorithm will be suitable and more accurate for prediction of the disease. DecisionmakingforpredictingthepresenceofdiabeticIrishisperformedusingK-Nearest Neighbor, Random Forest, Support Vector Machine and NeuralNetworks.

1. Introduction

Diabetes is a chronic and organ disease that occurs when the pancreas does not secrete enoughinsulinorthebodyisunabletoprocessitproperly.Overtime,diabetesaffectsthecircular system, including that of the irish. Diabetes Irish (DI) is a medical condition where the irish is damaged because of fluid leaks from blood vessels into the irish. It is one of the most common diabetic eye diseases and a leading cause of blindness. Nearly 415 million diabetic patients are at risk of having blindness because of diabetics. It occurs when diabetes damages the tinybloodvesselsinsidetheirish,thelightsensitivetissueatthebackoftheeye.Thistinyblood vessel will leak blood and fluid on the irish forms features such as micro-aneurysms, haemorrhages, hard exudates, cotton wool spots or venous loops. Diabetic Irish can be classified as non-proliferative diabetic Irish (NPDR) and proliferative diabetic Irish (PDR).Dependingonthepresenceoffeaturesontheirish,thestagesofDRcanbeidentified.In the NPDR stage, the disease can advance from mild, moderate to severe stage with various levels of features except less growth of new blood vessels. PDR is the advanced stage where the fluids sent by the irish for nourishment trigger the growth of new blood vessels. They grow along the irish and over the surface of the clear, vitreous gel that fills the inside of the eye. If they leak blood, severe vision loss and even blindness canresult.

Currently, detecting DI is a time-consuming and manual process that requires a trained clinician toexamineandevaluatedigitalcolourfundusphotographsoftheirish.Bythetimehumanreaders submit their reviews, often a day or two later, the delayed results lead to lost follow up, miscommunication, and delayedtreatment.

1.1 Objectives &Goals:

Thispapermainlyfocusesonthepredictionofdiabeticlrishandanalysisperformed of different algorithm for the prediction. Machine learning algorithms such as KNN, RF, SVM, NNET etc. can be trained by providing training datasets to them and then these algorithms can predictthedatabycomparingtheprovideddatawiththetrainingdatasets.Ourobjectiveis totrain our algorithm by providing training datasets to it and our goal is to detect diabetic Irish using different types of classificationalgorithms.

2. MachineLearning

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [4]. Machine learning algorithms use computational methods to“learn”informationdirectlyfromdatawithoutrelyingonapredeterminedequationasa model. The algorithms adaptively improve their performance as the number of samples available for learning increases. Tom M. Mitchell provided a widely quoted and more formaldefinition:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E [5].

The core of machine learning deals with representation and generalization. Representing the data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the ability of a machine learning system to perform accurately on new, unseen datainstancesafterhavingexperiencedalearningdatainstance.Thetrainingexamplescomefrom somegenerallyunknownprobabilitydistributionandthelearnerhastobuildageneralmodelabout this space that enables it to produce sufficiently accurate predictions in new cases. The performance of generalization is 11 usually evaluated with respect to the ability to reproduce known knowledge from newer examples. There are different types of machine learning, but the two main ones are:

- SupervisedLearning
- UnsupervisedLearning

3. Supervised LearningModel

Supervised learning is the machine learning task of inferring a function from supervised trainingdata[6].Trainingdataforsupervisedlearningincludesasetofexampleswithpairedinput subjects and desired output. A supervised learning algorithm analyses the training data and produces an inferred function, which is called classifier or a regression function. The function should predict the correct output value for any valid input object. This requires the learning algorithm to generalize from the training data to unseen situations in a reasonableway.

A simple analogy to supervised learning is the relationship between a student and

International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 03 | Mar -2020 ISSN: 2582-3930

a teacher. Initiallytheteacherteachesthestudentabout aparticulartopic.Teachingthestudenttheconcepts of the topic and then giving answers to many questions regarding the topic. Then the teacher sets an exam paper for the student to take, where the student answers newerquestions. Figure2.1describesthatthesystemlearnsfromthedataprovidedwhichcontainsthefeaturesand the output as well. After it has done learning, newer data is provided without outputs, and the systemgeneratestheoutputusingtheknowledgeitgainedfromthedataonwhichittrained. Here is how supervised learning modelworks.

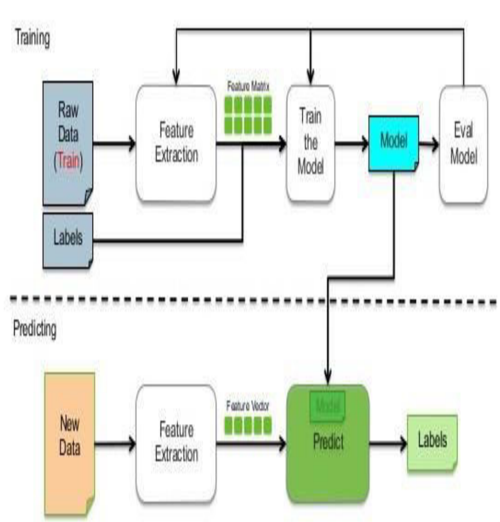


Figure 2.1: Workflow of supervised learning model

Algorithms

Since there are so many algorithms for machine learning, it is not possible to use all of themforanalysis.Forthisresearchpaper,we willbeusingfourofthemneuralnetworks(NNET), random forest (RF), K-Nearest Neighbor (KNN) and support vector machine(SVM).

4. NeuralNetworks

Within the field of machine learning n neural networks are a subset of algorithms built around a model of artificial neurons spread across three or more layers [7]. There are plenty of other machine learning model which is notable for being adaptive in nature. Every node of neural network has their own sphere of knowledge about rules and functionalities to develop it-self through experiences learned from previous techniques that don't rely on neural networks. Neural networksarewell-suitedtoidentifyingnon-linearpatterns,asinpatternswherethereisn'tadirect, one-to-one relationship between the input and output [8]. This is a learning training. Neural networksarecharacterizebycontainingadaptiveweightsalongpathsbetweenneuronsthatcanbe tunedbyalearningalgorithmthatlearnsfrom observeddatainordertoimprovemodel.On emust chooseanappropriatecostfunction.Thecost functioniswhatisusedtolearntheoptimalsol ution totheproblembeingsolved[7].Inanutshell,i tcanadjustitselftothechangingenvironmen tasit learns from initial training and subsequent runs provide more information about theworld.

5. RandomForest

Random forest algorithm can use both for classification and the regression kind of problems. It is supervised classification algorithm which creates the forest with a number of tress [9]. In general, the more trees in the forest the more robust the forest looks like. It could be also said that the higher the number of trees in the forest gives the high accuracy results. There are manyadvantagesofrandomforestalgorith ms.Theclassifiercanhandlethemissingval ues.Itcan also model the random forest classifier for categorical values [10]. The over fitting problem will never come when we use the random forest algorithm in any classification problem. Most importantly it can be used for feature engineering which means identifying the most important feature out of the available feature from the trainingdataset.

6. K-NearestNeighbors

K-nearestNeighborsisasimplealgorithmthats toresallavailablecasesandclassifiesnew cases based on a similarity measure [11]. KNN has been used in statistical estimation and pattern recognition.KNNmakespredictionforane winstance(x)bysearchingthroughtheentire training setforthekmostsimilarinstancesandsumm arizingtheoutputvariableforthosekinstanc es.For regression this might be the mean output variable, in classification this might be the mode class determine which of the k instances in the training dataset are most similar to new input many distance measure is used like Euclidean distance, Manhattan distance, Minkowskidistance.

7. Support VectorMachine

The Support Vector Machine (SVM) is a state-of-the-art classification method introduced in 1992 by Boser, Guyon, and Vapnik [12].

A more formal definition is that a support vector machine constructs a hyper plane or set ofhyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyper plane that has the largest distance to the nearest training data point of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier[13].

SVMs belong to the general category of kernel methods. A kernel method is an algorithm that depends on the data only through dot-products. When this is the case, the dot product can be replaced by a kernel function which computes a dot product in some possibly high dimensional featurespace.Thishastwoadvantages:First, theabilitytogeneratenon-lineardecisionboundaries using methods designed for linear classifiers. Second, the use of kernel functions allows the user toapplyaclassiertodatathathavenoobvious fixed-dimensionalvectorspacerepresentation[14].

8. PROPOSED MODEL FOR PREDICTION

This chapter contains proposed model, dataset collection, description, data visualization and also classifying algorithms that are used for analysis performance.

ProposedModel

Our First phase is data collection. We have collected our dataset from UCI Machine Learning repository website. The dataset contains features extracted from Messidor image set to predictwhetheranimagehavesignsofdiabeticIrishornot.Thenfeaturesandlabelsofthe dataset are identified. After that the dataset is divided into two sets, one for training where most ofthedataisusedandtheotheroneistesting.Intrainingsetfourdifferentclassificationalgorithms has been fitted for the analysis performance of the model. The algorithms we used are k-Nearest Neighbor, random forest, support vector machine and neural networks. After the system hasdone learningfromtrainingdatasets,newerdatais providedwithoutoutputs.Thefinalmodelgenerates the output using the knowledge it gained from the data on which it was trained. In final phase we get the accuracy of each algorithm and get to know which particular algorithm will give us more accurate results for the prediction of diabetic Irish.

Implementation

DataCollection

In our project we have used a dataset that is obtained from the UCI Machine Learning Repository. This dataset contains features extracted from Messidor image set to predict whether an image contains signs of diabetic Irish or not. All features represent either a detectedv lesion, a descriptive feature of an anatomical part or an image-level descriptor. The Messidor database has been established to facilitate studies on computer-assisted diagnoses of diabetic Irish. We have seen different kind of datasets in kaggle, github and other websites which was used for different kind of projects based on diabetic Irish. As we wanted to work with

detection of diabetic Irish, this dataset will be appropriate for our work as it has different types of features.

DataDescription

Our dataset contains different types of features that is extracted from the Messidor image set. This dataset is used to predict whether an image contains signs of diabetic Irish or not. The value here represents different point of irish of diabetic patients. First 19 columns in the datasetareindependentvariablesorinputcolumnandlastcolumnisdependentvariablesoroutput column.Outputsarerepresentedbybinarynumbers.“1”meansthepatienthasdiabeticIrish and “0” means absence of thedisease.

Feature indexes are-

i. q – The binary result of quality assessment. 0=bad quality 1= sufficientquality.
ii. ps –The binary result of pre-screening, where 1 indicates severe irishl abnormality and 0 its lack.
iii. nma.a - nma.f - The results of microaneurism detection. Each feature value stand for the number of microaneurisms found at the confidence levels alpha = 0.5, . . . , 1,respectively.
iv. nex.a – nex.h - contains the same information as nma.a - nma.f for exudates. However, as exudates are represented by a set of points rather than the number of pixels constructingthe lesions, these features are normalized by dividing the number of lesions with the diameter

of the ROI to compensate different imagesizes.

v. dd - The euclidean distance of the center of the macula and the center of the optic disc to provide important information regarding the patient's condition. This feature is also normalized with the diameter of theROI.

	nex.h	dd	dm	amfm	class
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	0.037225	0.523212	0.108431	0.336229	0.530843
std	0.178959	0.028055	0.017945	0.472624	0.499265
min	0.000000	0.367762	0.057906	0.000000	0.000000
25%	0.000000	0.502855	0.095799	0.000000	0.000000
50%	0.000000	0.523308	0.106623	0.000000	1.000000
75%	0.003851	0.543670	0.119591	1.000000	1.000000
max	3.086753	0.592217	0.219199	1.000000	1.000000

vi. dm-The diameter of the opticdisc.

vii. amfm - The binary result of the AM/FM-basedclassification.

viii. class - Class label. 1 = contains signs of Diabetic Irish, 0 = no signs of Diabetic Irish.

We have also calculated count, mean, max, standard deviation of the values in our dataset.

	q	ps	nma.a	nma.b	nma.c
count	1151.000000	1151.000000	1151.000000	1151.000000	1151.000000
mean	0.996525	0.918332	38.428323	36.909644	35.140747
std	0.058874	0.273977	25.620913	24.105612	22.805400
min	0.000000	0.000000	1.000000	1.000000	1.000000
25%	1.000000	1.000000	16.000000	16.000000	15.000000
50%	1.000000	1.000000	35.000000	35.000000	32.000000
75%	1.000000	1.000000	55.000000	53.000000	51.000000
max	1.000000	1.000000	151.000000	132.000000	120.000000

DataVisualization

Another important feature in the data distribution is the skewness of each class. Data visualization helps to see how the data looks like and also what kind of data correlation we have. The dataset distribution of each feature is shown below in figure 3.5. This is a histogram. Ahistogram is an accurate graphical representation of the distribution of numerical data. It is anestimateoftheprobabilitydistributionofa continuousvariable.Histogramsareagreat waytoget to know your data. They allow you to easily see where a large and a little amount of the data can be found. In short, the histogram consists of an x-axis and a y-axis, where the y-axis shows how frequently the values on the x-axis occur in thedata.

As the given input variables are numeric, we can also create boxplot.

A Boxplot typically provides the median, 25th and 75th percentile, min/max that is not an outlier and explicitly separates the points that are consideredoutliers.

SplitDataset

Separatingdataintotrainingandtestingsetsisanimportantpartofevaluatingdata mining models. Typically, when separating a data set into two parts, most of the data is used for training, and a smaller portion of the data is used for testing. We have also split our dataset into two sets. Oneisfortrainingandanotherfortesting.Th etrainingsetcontainsaknownoutputandthe model learns on this data in order to be generalized to other data later on. After the model has been processed by using the training set, we have tested the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct or not. Inaddition, we have used 80% of our data for training and 20% fortesting.

ApplyingAlgorithm

We went through a process of trial and error to settle on a short list of algorithms that provides better result as we are working on classification of diabetic Irish, we used some machinelearningclassificationalgorithms. Wegetanideafromthedatavisualizationspl otswhich algorithms will be suitable for the classification problem. The Machine Learning system uses the trainingdatatotrainmodelstoseepatterns,a ndusesthetestdatatoevaluatethepredictive quality of the trained model. Machine learning system evaluates predictive performance by comparing predictions on the evaluation data set with true values (known as ground truth) using a variety of metrics.

So, for our thesis we will evaluate four different machine learning algorithms –

- Neural Networks(NNET)
- RandomForest
- K-Nearest Neighbor(KNN)
- Support Vector Machine(SVM)

K-Fold CrossValidation

K-Fold Cross Validation is common types of cross validation that is widely used in machine learning. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k-1 subsamples are used as training data. In our project we used 10-fold cross validation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

SystemSetup

Hardware and software used in this research played a big role in terms of results. Both hardware and software specifications have been mentioned here.

HardwareSpecification
SoftwareSpecification

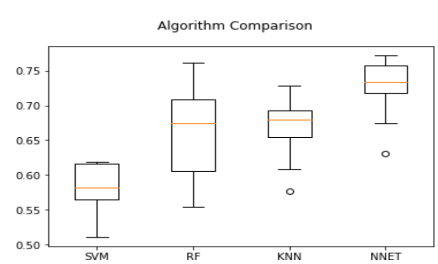
EXPERIMENTAL RESULTS & ANALYSIS

In the previous chapter we have discussed about proposed system and implementation of our thesis. We have demonstrated how we collected our dataset, dataset description, visualization and algorithms we used. Now we discussing about the results we obtained from our experiments upon the implementation of this system. We have divided our dataset into two parts- training and testing dataset. In this chapter we will show the outcome of the training and testing dataset. As mentioned before we have used four machine learning algorithms. First, we trained our dataset with these four algorithms and then we built a model. Then, we tested our testing dataset in this model. If the test set accuracy is near to train set accuracy then we can conclude that we built a good model. Wehavetotal1151dataofdifferentindividualinourdataset.Thereare1151rowsand20columns in the dataset. After splitting the data into two parts now we have 920 rows for train data and for test data we have 231 rows. When we trained our train data for analysis performance of different algorithms. This is the result wegot-

Comparison betweenAlgorithms

A comparison between the algorithms we used for our training dataset. Here,thetalllineindicatesstandarddeviationandtherectangularboxindicatesmedianvalueand thebrownlineintheboxindicatesthemeanvalue.Fromherewecanunderstandwhichalgorithm is good for ourmodel.

Figure 4.5: Comparison between algorithms



After training the model we test the model with the testing dataset. We have 20% data for testing inthetestingset.Table4.1showsthetestinga ccuracy,precision,recallandF1score.Thed etailed information of the test data evaluation with unigram model is asfollows-

Table 4.1: Accuracy of test dataset

Models	Accuracy	Precision	Recall	F1 Score
SVM	57.07 %	62%	57%	53%
KNN	64.50 %	65%	65%	65%
RF	63.63 %	64%	64%	64%
NNET	75.32 %	78%	75%	75%

In experimental result, we observe that the accuracy of the both training and testing set is quite similar and for both training and testing dataset NNET algorithm is giving higher accuracy rate which is around 75%.So, we can say that this algorithm will give us more accurate prediction about the disease. As our main purpose of the thesis is to build a model which will classify the diabetic Irish as accurate as possible, we hope that this final model will give us proper and appropriate results.

We have also determined our train and test model accuracy and loss. For this visualization model wehaveusedkeraspackageforobtainingthis trainandtest-lossandaccuracy.Wehavealsoused Historycallbackforthispurpose.Oneofthed efaultcallbacksthatareregisteredwhentrain ingall deep learning models is the History callback. It records training metrics for each epoch. This includesthelossandtheaccuracy(forclassifi cationproblems)aswellasthelossandaccur acyfor the test dataset, if one isset.

The history object is returned from calls to the fit function used to train the model. Metrics are stored in a dictionary in the history member of the object returned.

CONCLUSION

This chapter contains the difficulties, future works and concluding remarks, which will give the summary of our thesis work and also give the indication of our future plan with our thesis project.

References

[1] Gandhi M. and Dhanasekaran R. (2013). Diagnosis of Diabetic IrishUsing Morphological

Process and SVM Classifier, IEEE International conference on Communication and Signal Processing, India pp:873-877

[2] Li T, Meindert N, Reinhardt JM, Garvin MK, Abramoff MD (2013) Splat Feature ClassificationwithApplicationtoIrishlHemorrhageDetectioninFundusImages,IEEE Transactions on Medical Imaging, 32:364-375

[3] Yau JW, Rogers SL, Kawasaki R, Lamoureux EL, Kowalski JW, Bek T, et al. Global prevalence and major risk factors of diabetic Irish. Diabetes Care.2012;35:556–64

[4] BoserB ,Guyon I.G,Vapnik V., "A Training Algorithm for Optimal Margin Classifiers", Proc. Fifth Ann. Workshop Computational Learning Theory,pp. 144-152, 1992.

[5] Mitchell, T. (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7.,McGraw-Hill, Inc. New York, NY, USA. Published on March 1, 1997

[6] Alex C, Boston A. (2016).Artificial Intelligence, Deep Learning, and NeuralNetworks, Explained(16:n37)

[7] Saimadhu P. How the Random Forest Algorithm Works in Machine Learning. Published on May 22, 2017

[8] Boulesteix, A.-L., Janitza, S., Kruppa, J., & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics.WileyInterdisciplinaryReviews:DataMiningandKnowledgeDiscovery, 2(6),493–507.

[9] Jason B, Boinee P.“Machine Learning Algorithms”2(3), 138–147. Published on 15,2016.

[10] Boser B. E, Guyon I. M.,Vapnik V. N. (1992). “A training algorithm for optimal margin classiers”.Proceedings of the 5th Annual Workshop on Computational Learning Theory COLT'92, 152 Pittsburgh, PA, USA. ACM Press, July 1992. On Page(s):144-152