ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse43922020.pdf
https://doi.org/10.30534/ijatcse/2020/43922020

Diabetic Prediction Using Kernel Based Support Vector Machine

Annavarapu Naga Prathyusha [1] M.R.Narasinga Rao [2]
[1] M.TECH Student, Koneru Lakshmaih Educational Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India. annavarapuprathyusha@gmail.com
[2] Professor, Koneru Lakshmaih Educational Foundation (KLEF), Vaddeswaram, Guntur District, Andhra Pradesh-522502, India.ramanarasingarao@kluniversity.in

ABSTRACT

Insulin Dependent Diabetes is metabolic (total chemical reactions occurring in the body) disorder which is characterised by high glucose levels (hyperglycemia), glycosuria (glucose in urine), hyperlipidemia (high lipid levels).According to world statistics for prevalence of diabetes, there are 415 million people in the world are suffering from diabetes and it was predicted that, out of 10 in 100 of worlds adult population affected with diabetes. Also, it is observed that, 46% of people with diabetes are not diagnosed. By the year of 2040, it is expected that the count may rise to 642 million all over the globe. In this proposed research a classifier is designed by the use of machine learning techniques, to predict the disease given the inputs of the patient. The dataset used is Pima Indian Diabetes which is used for prediction of diabetes. The proposed system use the support vector machine algorithm by using distinct types of kernels such as sigmoid kernel, radial basis kernel (rbf), polynomial kernel, linear kernel, and also applied normalization by using standard scaler(sklearn library) with visualizations before and after normalizations. Based on this results generated by each kernel using SVM we compared the performance of the model with different kernels and found that the proposed model could be used to predict the disease accurately.

Key words: SVM, kernels, diabetes, normalization, standard scaler.

1. INTRODUCTION

Diabetes Mellitus (DM) is laid out as a variety of metabolic disorders basically brought about by unusual insulin emission and additionally activity [3].Insulin helps keeps your glucose level from getting upscale (hyperglycemia) or downscale (hypoglycemia) [1].The start of diabetes is surveyed to rise fundamentally in the following years [1].There are two essential divisions, type 1 diabetes followed by type 2 diabetes as demonstrated by the etiopathology of the turmoil. Type 2 diabetes appears, apparently, to be the most broadly perceived sort of diabetes mainly depicted by insulin hindrance [3].

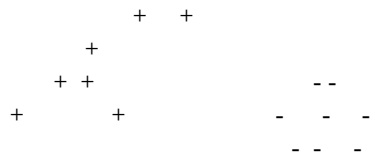
1.1Supervised Learning

In Machine Learning Supervised Learning can be simply defined as it learns from the experience from the input labels (which can be known). There are two types of tasks we can perform using machine learning techniques and they are Classification and Regression, where classification is to predict discrete values having a place with a specific class and assess based on accuracy. It is very well may be either binary or multi class classification. In paired order, model predicts either 0 or 1; yes or no however if there should be an occurrence of multi class arrangement, model predicts more than one class. Example, Gmail has its classification into more than three classes such as social, followed by promotions, followed by updates, and finally with forums. The intent of regression is to predict output having continuous value of dependent variable and an example for which is Wind Speed which doesn't have any discrete value but is continuous in the particular range. [3, 5].

1.2 Support Vector Machine (SVM)

SVM was evoked by Vladimir N.Vapnik in the year 1963. At the time of Bernhard E. Boser, followed by Isabelle M. Guyon and also Vladimir N. Vapnik ordered a way to build nonlinear classifiers by applying the kernel trick to maximum-margin hyper planes**.**[1,5]. SVM is distinctive classification model among all other with good generalization ability on unseen data [2] and also it is good at performing tasks like classification and regression [7].Not only the svm comes under supervised learning, but also to minimize the classification error and maximize the geometric margin. This is the reason why SVM is also called as Maximum Margin Classifier [7].

Let us assume that we want to separate '+' and '-'in the following diagram:



Let us assume that there are training examples such as (x_1, y_1)……(x_n,y_n) [here these terms are called predefined labels]

For each Example i: $x_i = (x_i^{(1)}, \ldots\ldots\ldots\ldots , x_i^{(d)})$ here x_i is the feature vector and $x_i^{(1)}$, $x_i^{(d)}$ defines whether the values are present or not

$X_i^{(j)}$ is real valued

Y_i belongs to {-1, +1}, y_i is a class label whether it is Yes or No

Now for Classification, the Inner product can be defined as product of Weight vector W with Feature vector X to form classification

$W \cdot X = \sum_{j=1}^{(d)} W^{(j)} \cdot X^{(j)}$

1.3 Kernel Functions

SVM algorithms use a set of mathematical functions that are defined as the kernel trick. If the problem is non-linear, instead of trying to fit a nonlinear model from the original space, we try to map the problem to a new space by doing a nonlinear transformation using suitably chosen basis function and then use a linear model in the new space. The linear model in the new space is equal to fitting the non-linear model from the required original space. This we call kernel trick and hence can be used in solving nonlinear problems in different domains. The motive behind the usage of kernel function is mapping from lower dimension to higher dimension [2]. These functions can be different types. For example linear kernel, nonlinear, polynomial, radial basis function followed by sigmoid [2, 5].The detailed Explanation of this paper is followed by literature survey, methodology, results, conclusion and references.

2. RELATED WORKS

The authors Dr. R. Vijayakumar, Kavin Prasad Arjunan, Manivel Sivasakthi, and Karthikeyan Lakshmanan have proposed a framework on prediction of diabetes by using support vector machine and k-means clustering. They have implemented the project by using .net framework i.e., windows/desktop framework with a backend to support for storing the database. The advantage of this proposed system is that, it extracts options mechanically for structured and unstructured information, and it provides correct attainable risk predictions. It combines the structured and unstructured information to assess the likelihood in prevalence of polygenic disease [1].

The authors Deepika Kancherla, Jyostna Devi Bodapati, Veeranjaneyulu N have defined different types of kernels for finding the performance of support vector machine classification by using different types of datasets such as face recognition, scene recognition, tumour detection, MIT dataset, MNIST dataset and digit recognition by applying linear kernel, radial basis kernel, polynomial kernel and sigmoid kernel for the datasets. They also came up with kernel gram matrix which defines and helps to find out the kernels that best classifies the data and also visualizations are done in this project and given the future scope is to work with deep neural networks [2].

The authors Tejas N. Joshi1, Prof. Pramila M. Chawan have predicted the diabetes based on logistic regression. Logistic Regression is mainly good for binary classification.

The logistic function is defined as:

Transformed = 1 / (1 + e^-x)

Where e stands for numerical constant Euler's number and x is an input we are giving into the function. The authors have proved that, by using the logistic regression, they have come up with 78% accuracy [3].

The author Abdul Azis Abdillah, Suwarno have used radial basis kernel, which can be called as RBF (in short) for the prediction of diabetes. Diabetes dataset has been used for the purpose. The model is trained and tested with Ten-fold cross validation and concluded by checking the performance of classifier by using confusion matrix and ROC curve with an accuracy of 80.22% [4].

The authors Supriya Pahwa, Deepak Sinwar used four different types of kernels i.e., linear kernel, radial basis kernel, polynomial kernel and sigmoid kernel and compared all the kernels by using different types of datasets they are a1a, a6a, w7a and Australia and also calculated the prediction time and calculation of trained time for the 4 types of kernels and also compared the cpu time of four different kernels and finally described that among all the kernels they have mentioned that linear kernel gave the improved performance than the remaining three kernels with an accuracy of 88.2% and prediction time of accuracy 4.078 seconds. Though the linear kernel gave the best performance, the radial basis kernel took less time than other kernels [5].

The authors B. Yekkehkhany, A.Safari, S. Homayouni, and M.Hasanlou have developed a system using Support Vector Machines (SVM) for crop grouping utilizing polarimetric highlights separated from multi-transient Synthetic Aperture Radar (SAR) symbolisms. The multi-fleeting joining of information improves the general recovery exactness as well as furnishes progressively dependable appraisals concerning single-date information. Very few piece capacities are utilized and thought about in this investigation for mapping the info space to higher Hilbert measurement space. These piece capacities incorporate straight, polynomials and Radial Based Function (RBF). The technique is applied to a few UAVSAR L-band SAR pictures obtained over a horticultural territory close to Winnipeg, Manitoba, Canada. In this exploration, the transient alpha highlights of H/A/α disintegration strategy are utilized in grouping. The trial tests show a SVM classifier with RBF portion for three dates of information builds the Overall Accuracy (OA) to up to 3% in contrast with utilizing direct piece work, and up to 1% in contrast with a third degree polynomial bit work[6].

The authors V. Anuja Kumari1, R.Chitra have developed a model and applied a classification algorithm using support vector machine popularly called as SVM and applied this algorithm for diabetes dataset and generated the results by using confusion matrix and expressed in ROC(Receiver Operating Characteristic) Curve[7].

The authors Rahul Samant, Srikantha Rao have identified performance of different models using Support Vector Machines (SVMs) using different kernel functions to predict the chance of person who can be proned to be diabetic or not. The SVM Algorithm has been trained with 13 inputs from the medical dataset. Different kernel functions, like Linear, followed by Quadratic, followed by Polyorder, followed by Multi-Layer Perceptron and Radial Basis kernel were coded and tested to create the diagnosing system. All the kernels were applied for the dataset and showed a fairly smart accuracy for linear kernel [8].

The Authors D. Ben Ayed Mezghani, S. Zribi Boujelbene, N. Ellouze had one of the central issues with the study of SVM is the kernel choice and that's primarily based on the matter of selecting a kernel operator for a specific task and dataset and Compared the performance to several Machine Learning algorithms. The author had a kernel choice of SVM classifier to realize performance on text-independent recognition mistreatment for the TIMIT corpus. In this study, they have used SVM classifier linear, polynomial and Radial Basis Function were used for the purpose. A study has been created between SVM mistreatment with the most effective selection of kernel and 3 other different learning algorithms viz., specifically Naive Bayes, call tree C-4.5 and MLP have been used for the purpose. Results have shown that, SVM with polynomial kernel is the best option for addressing recognition tasks in comparison to different algorithms. [9]. S.Uddaraju and M R Narasingarao have developed machine learning techniques in predicting the ductal carcinoma [10] Similar work has been done by T.Sajana and M R Narasingarao in classification of malaria disease[11]. A comparative study of SVM and Logistic Regression has been done by Deepthi Gurram and M R Narasingarao for the diagnosis of thyroid disfunction [12]. Mana Ranjan Senapathi have developed detection and classification methods for EEG epileptic seizures [13]. Similarly, A. Nagarajan and J.vasanth Wason have used a Machine Learning approach to predict Lung Cancer using CT Scan Images [14]. Venubabu Rachapudi, Krishna, Sai, S.Hari priya and K.Pushpahas have developed an effective approach to classify retinal images for diabetic retinopathy [15]

3. APPROACHES OF DIFFERENT KERNELS

3.1 Comparison Of Different Kernels

In this paper, the dataset used is Pima Indian diabetes on four different types of kernels. They are sigmoid kernel, linear kernel, polynomial kernel and radial basis kernel and these are all compared with Support Vector Machine before and after applying normalization.

Linear Kernel is commonly used for linear data, and for this, we use a factor named regularization parameter(c=1.0) which makes the kernel faster to execute, and where as in other types of kernels, we use ϒ parameter meant for grid search.

Radial Basis Kernels are more expensive when compared to linear kernels and this kernel is generally preferred when the data is non-linear. We use a simple factor called ϒ parameter, which is used only in radial basis kernel. As the value of ϒ increases, then the model gets overfit and if the value of ϒ decreases then the model gets underfit.

The Polynomial Kernel is non-linear in nature and to compute this kernel, it is very complex to compute as the power increases the complexity.

$$K(x1, x2) = (x1. x2 + 1)^{d}$$

Where, d is the degree of the polynomial and x1 and x2 are vectors

Let's assume X space (Sample Data) is 2-dimensional so that

$$Px = (r1, r2)$$

$$Py = (s1, s2),$$

Here P_x, P_y are 2-Dimensional data where r1, r2 and s1, s2 are values

So, if we map this into higher dimension for our convenience, let us assume in Q-space where it is six-dimensional the result will be

$$\mathbf{Q_a=\Phi(Px)=(1,r1,r2,r1^2,r2^2,r1*r2)}$$

$$\mathbf{Q_b=\Phi(Py)=(1,s1,s2,s1^2,s2^2,s1*s2)}$$

Sigmoid kernel and Tangent hyperbolic kernel are normally used in Multi-layer Perceptron Neural Networks for knowing the behaviour of the network. Normalization is data pre-processing techniques and the target of normalization is to change the values of columns to equivalent scale so that it should not change any differences in the range of values.

3.2 Proposed System

In this proposed research, we have developed a system for prediction of diabetes and the Architecture of the proposed system is as shown in Figure 1.

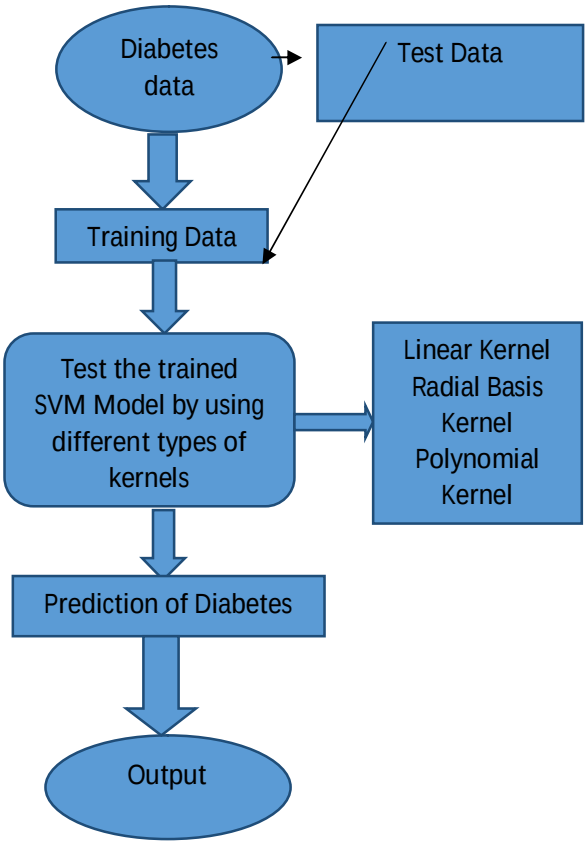


Figure 1: Architecture for Proposed System

3.3 Methodology

3.3.1 Preprocessing: Convert the given numerical data into non-missed values without any missing values if the data contains any missing values then those missing values should be removed, else it leads to wrong prediction and less accuracy.

3.3.2 Kernels Function: SVM models are developed using the Linear Kernel, Radial Basis Kernel Function, Polynomial Kernel and Sigmoid Kernel. The Kernel trick is given by the following equation:

$$K(\vec{X}i,\vec{X}j)=\Phi(\vec{X}i)^T\,\Phi(\vec{X}j)$$

The Following are the most used kernels and their equations are:

$K(X_i,X_j)=(X_i.X_j+a)^P$ where P is the degree of kernel and a stands for constant

$K(X_i, X_j)=\exp(-\Upsilon||x_i-x_j||^2)$ where Υ is kernel parameter followed by x_i is training vector

$K(X,X_i)=\text{sum}(x*x_i)$ gives the Linear Kernel

$K(X,X_i)=\tanh(\eta.X_i.X_j+v)$---Tangent Hyperbolic kernel Activation function for neural network

3.3.3 Model Selection: Support Vector Machine Algorithm is used and various types of kernels are selected to find the best accuracies.

3.3.4 Train and Test: The models have been trained and tested for all the four different types of kernels used in the diabetes dataset.

3.3.5 Dataset: The dataset used in this paper is Diabetes Data which consists of 768 rows and 8 columns and with one target/outcome variable which includes 0 and 1 where 0 indicates not diabetic and 1 indicates diabetic. As there were no missing values, we included all the rows and columns. This dataset is taken from UCI repository which is also available in Kaggle Repository. Each attribute in the dataset describes as follows:

1) Pregnancies describes the number of times the person has been pregnant.
2) Glucose describes the blood glucose level on testing.
3) Blood pressure describes the diastolic blood pressure.
4) Skin Thickenss describes the skin fold thickness of the triceps.
5) Insulin describes the amount of insulin in a 2-hour serum test.
6) BMI describes the body mass index.
7) DiabetesPedigreeFunction describes the family history of the person.
8) Age describes the age of person
9) Outcome describes if the person is predicted to have diabetes or not.

4. RESULTS

The accuracies and visualizations are computed and the results are shown in the following Table 1.

Dataset	Accuracy for linear kernel	Accuracy for radial basis kernel	Accuracy for polynomial kernel	Accuracy for sigmoid kernel
Diabetes	80.5	64.9	62.3	64.9
Diabetes(After applying Normalization using sklearn library as StandardScaler)	80.0	76.6	71.4	76.6

Table 1: Accuracies for different types of kernels

In the following the graph, we represent the different types of kernels such as Linear Kernel, Radial Basis Kernel, Polynomial Kernel and Sigmoid Kernel. Among all these types of kernels, Linear Kernel has given an accuracy of 80.5% followed by Radial Basis kernel with an accuracy of 65%, followed by Polynomial Kernel (degree=8) with an accuracy of 62.3% and finally the Sigmoidal Kernel gave an accuracy of 64.9%.In this, we have compared with two tasks and the objective is to enhance the accuracy of different types of kernels after applying standard scaler (sklearn library).

4.1 Histogram for diabetes dataset

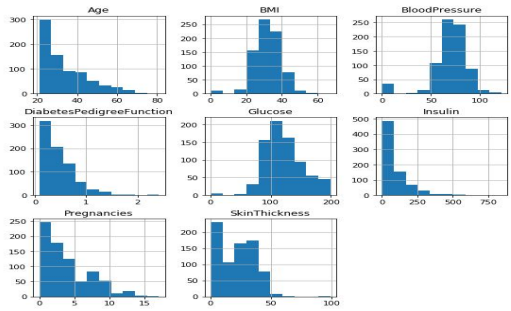


Figure 2: Architecture for histogram

A Histogram is one of the best Data visualization Techniques popularly known in Machine learning. It means the division of a continuous variable over a period of time or over a given interval. Histogram divides the intervals which are called as bins. Each of the distribution represents normal distribution, skewness and outliers etc.

The Figure 2 Histogram Distribution clearly shows the different attributes in diabetes data such as Age, BMI, Blood Pressure, Diabetes Pedigree Function, Glucose, Insulin, Pregencies, and Skin Thickness.

The Inference for Age shows the Clear Distribution such that X-Axis represents the Range of the Age in the diabetes dataset (For example the range for the age is from 20-80) and Y-Axis represents the count of the Age (For example in our dataset people with age 60-80 are nearly 31) and also the Age plot represents that it is skewed. Skewness can be quantified to define that which distribution differs from normal distribution.

The Inference for BMI shows the Clear Distribution such that X-Axis represents the Range of BMI in the diabetes and Y-Axis represents the count of the BMI and also the BMI plot represents that some people are Obese. Obese people have greater chance of developing diabetes.

The Inference for Blood Pressure shows the Clear Distribution such that X-Axis represents the Range of the BP in the diabetes dataset and Y-Axis represents the count of the BP and also the Blood Pressure plot looks like Normal Distribution with a normal diastolic values of 80.

The Inference for Diabetes Pedigree Function shows the Clear Distribution such that X-Axis represents the Range of the Diabetes Pedigree Function in the diabetes dataset and Y-Axis represents the count of the Diabetes Pedigree Function.

The Inference for Glucose shows the Clear Distribution such that X-Axis represents the Range of the glucose in the diabetes dataset and Y-Axis represents the count of the Glucose. From the glucose we can clearly observe that people who are diabetic have more glucose levels.

The Inference for Insulin shows the Clear Distribution such that X-Axis represents the Range of Insulin in the diabetes dataset and Y-Axis represents the count of Insulin. Most of the people who have high Insulin Values are also prone to Diabetic.

The Inference for Pregencies shows the Clear Distribution such that X-Axis represents the Range of the Age in the diabetes dataset and Y-Axis represents the count of Pregencies in the dataset.

The Inference for Skin Thickness shows the Clear Distribution such that X-Axis represents the Range of the Skin Thickness in the diabetes dataset and Y-Axis represents the count of the Skin Thickness and also the Skin thickness plot represents that it is skewed.

4.2 Correlation plot for diabetes dataset

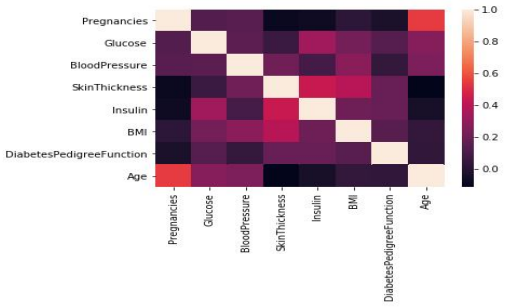


Figure 3: Architecture for Correlation Plot

Correlation are very helpful for understanding relationship between the two continuous variables and also it describes the how both the variables are varying in their direction either in same or opposite direction, if it goes in same direction then it is called Positive Correlation and if it goes in opposite direction then it is called Negative Correlation.

For the above correlation Plot shown in Figure 3 describes the diagonal part of the plot represents that for attribute in the diabetes Dataset does not exceed the 1.0. Generally the normal value for correlation is 1.0

4.3 Final Plots for Before and after Normalization

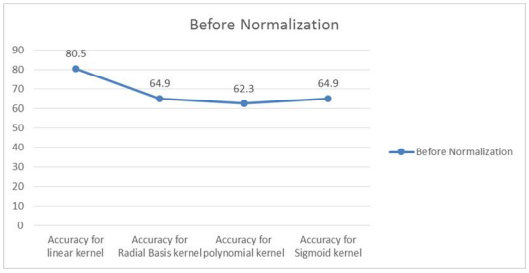


Figure 4: Accuracies for Kernels for different types of kernels

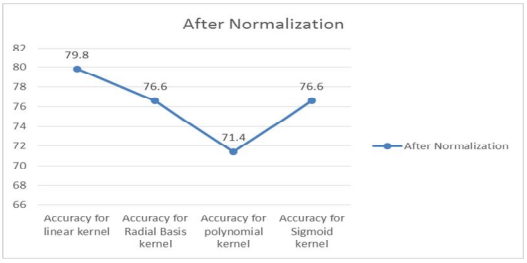


Figure 5: Accuracies for Kernels for different types of kernels after normalization

The above graphs (Figure 4, Figure 5) represent the before and after normalization where after applying normalization the accuracies are increased.

5. DISCUSSION

Normalization plays an essential role in Machine learning mostly for when applying to datasets. By importing StandardScaler it standardizes the data without changing the shape of data. Each and every column/feature in the data will have mean= 0 and standard variance = 1.Generally if the data is reduced then it is helpful for storage as well as it also helps in redundancy and also useful for normal forms such as 1-NF, 2-NF, etc.

6. CONCLUSION

In this paper, we have built different models using Support Vector Machine with distinct types of kernels like radial basis kernel, linear kernel, sigmoid kernel and polynomial kernel and performance of the model has been performed. It is observed that, linear kernel has performed better when compared with other kernels when it comes to prediction of the disease. These models has been trained and tested and found that SVM is a better model in the prediction of diabetes using different kinds of kernels.

REFERENCES

[1] Dr. R. Vijayakumar, Kavin Prasad Arjunan, Manivel Sivasakthi, Karthikeyan Lakshmanan, **Diabetes Prediction By Machine Learning Over Big Data From Healthcare Communities**, International Research Journal of Engineering and Technology (IRJET), Vo: 06 Issue: 04 | Apr 2019

[2] Deepika Kancherla, Jyostna Devi Bodapati, Veeranjaneyulu N, **Effect of Different Kernels on the Performance of an SVM Based Classification,** International Journal of Recent Technology and Engineering (IJRTE), Vol:7, Issue-5S4, February 2019

[3] Tejas N. Joshi, Prof. Pramila M. Chawan, **Logistic Regression and Svm Based Diabetes Prediction System**, International Journal for Technological Research In Engineering Vol: 5, Issue 11, July-2018

[4] Abdul Azis Abdillah, Suwarno**, Diagnosis of diabetes using support vector machines with radial basis function kernels,** International Journal of Technology, 2016
https://doi.org/10.14716/ijtech.v7i5.1370

[5] Supriya Pahwa, Deepak Sinwar, **Comparison of Various Kernels of Support Vector Machine**, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol: 3 Issue VII, July 2015.

[6] B. Yekkehkhany, A. Safari, S. Homayouni, M. Hasanlou, **A Comparison Study of Different Kernel Functions for SVM-based Classification of Multi-temporal Polarimetry SAR Data**, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol: XL-2/W3, 2014.

[7] V. Anuja Kumari, R.Chitra, **Classification Of Diabetes Disease Using Support Vector Machine**, International Journal of Engineering Research and Applications (IJERA), Vol. 3, Issue 2, March -April 2013

[8] Rahul Samant, Srikantha Rao, **Performance of SVM Classifiers in Predicting Diabetes**, International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 12, December – 2013

[9] D. Ben Ayed Mezghani, S. Zribi Boujelbene, N. Ellouze, **Evaluation of SVM Kernels and Conventional Machine Learning Algorithms for Speaker Identification**, International Journal of Hybrid Information Technology, Vol: 3, July-2010

[10] S.Uddaraju, M R Narasingarao, **Predicting the ductal carcinoma using machine learning techniques-A Comparison**, Journal of Computational and Theoretical Nano science, 16(5-6), 1902-07, 2019
https://doi.org/10.1166/jctn.2019.7822

[11] T.Sajana, M R Narasingarao, **An Ensemble Framework for Classification of Malaria Disease**, ARPN Journal of Engineering and Applied Sciences, 13(9), 3299-3307, 2018

[12] Deepthi Gurram, M R Narasingarao, **A Comparative study of SVM and Logistic Regression for the diagnosis of thyroid disfunction**, International Journal of Engineering & Technology, 7(1.1), 326-28, 2018.
https://doi.org/10.14419/ijet.v7i1.1.9714

[13] Sreeleka Panda, Satya Sis Mishra, Manas Ranjan Senapathy, **Detection and Classification Methods for EEG Epiletic Seizures**, International Journal of Advanced Trends in Computer Science and Engineering, Vol:8, No:6, pg.No:2925-2934, Nov-Dec,2019.
https://doi.org/10.30534/ijatcse/2019/40862019

[14] A. Nagarajan, J.vasanth Wason, **Machine learning approach to predict Lung Cancer using CT Scan images**, International Journal of Advanced Trends in Computer Science and Engineering, Vol:8, No:6,Pg.No: 2925-2934, Nov-Dec, 2019.
https://doi.org/10.30534/ijatcse/2019/48862019

[15] Venubabu Rachapudi, T.Krishna Sai, S.Hari priya, K.Pushpahas, **An effective approach to classify Retina Images for Diabetic retinopathy**, International Journal of Advanced Trends in Computer Science and Engineering Vol:8, No:6, Pg.No:2925-2934, Nov-Dec, 2019
https://doi.org/10.30534/ijatcse/2019/106862019