Related work

One of the worldwide increasing diseases is diabetes and the disease will subsequently get increasing if countries do not consider preventive measures for disease. Full Analysis and consideration of a patient are not possible nowadays by a doctor as various diseases are so widespread. And so an intelligent system is required for to consider the various factors and identify a suitable model between the different parameters. The use of intelligent systems in the diagnosis and treatment of diseases can significantly reduce medical errors leading to a decrease in both financial and human losses. The aim of this study is to compare the performance of several different classification techniques on a set of data that has been screened for diabetes.

The upcoming most popular and important approach in the field of medical research is Machine Learning. The frequency of diabetes over the years has been increasing dramatically with the aging population worldwide. Increased mortality rate and reduction in the life expectancy of elderly diabetic patients are due to growing diabetes. In clinical diagnosis problems, classification plays a vital role in further treatment of the disease. Various studies have been done on the diabetes data classification using different machine learning algorithm. In this Literature review section we are going to discuss work done on all the classification of diabetes using intelligence system, which are as follows:

**Support Vector Machines**

Nahla H *et.al.* (2010) have discussed a hybrid model for medical diagnosis which integrates three different data mining and SVM. They highlighted the data mining and machine learning techniques for diagnosis, prognosis, and management of diabetes.  SVMs is employed for finding a linear hyperplane that separates the positive and negative examples with an utmost interclass distance. However, **SVM** does not provide intelligible support for the classification decisions because it is behaving like black-box models. Hence, authors presented idea of intelligible representation of the SVM’s by using data extraction so that diagnosis can be performed efficiently. The penitent dataset was taken for this work from Oman. Results on the real-life prediction of type2 diabetes dataset show a comprehensible rule set which means intelligible SVM provides a promising tool for the prediction of diabetes with prediction accuracy of 94%, sensitivity of 93%, and a specificity of 94% [3][4].

*Kemal Polat* et.al. The aim of this paper is diagnosis of diabetes disease and so a new cascade learning system is been proposed that has 2 stages. The first stage contains Generalized Discriminant Analysis(GDA)[1] to discriminant feature variables between diabetic patients and healthy data as pre-processing process. The second stage uses LS-SVM[2] in order to classify diabetes dataset. The dataset used was Pima Indians diabetes dataset was selected from a larger data set held by the National Institutes of Diabetes and Digestive and Kidney Diseases. The result obtained when using LS-SVM is 78.21% classification accuracy using 10-fold cross validation, the GDA–LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation. The highest accuracy obtained was 82.05% and is very promising with regards to other classification. The fallout strongly suggest that Generalized Discriminant Analysis and Least Square Support Vector Machine classifier are based on a learning method can assist in the diagnosis of Diabetes disease.

*Seokho Kang* (2015)In this study, we propose an effective and efficient ensemble of SVMs for large-scale datasets based on data collection methods, called E3 –SVM. SVM [2] is one of the most accepted state-of-the-art classification algorithms, and based on the structural risk minimization principle it shows superior generalization performance. The projected method utilizes data selection methods to lessen the training set before constructing a collection. The projected method achieved comparable success in less time. Also, more reliable results are obtained for each independent run of constructing a collection. The model for the prescription level calculation derived by our projected method achieved about 80% classification accuracy for prescription cases.for future scope of improvement is thereby improving the effective data selection method.

*T. Santhanam , M.S Padmavathi* It is very difficult to handle the vast amount of data in medical and so applications of medical data include artificial neural network, fuzzy system, genetic algorithms, rough set, and support vector machine [1] which helps to reduce the difficulty. One of the major research areas where clustering algorithms and evolutionary algorithms play a vital role is Medical Data Mining. And one important application of medical mining is genetic algorithms. This research works on  Genetic Algorithms which are used for finding the optimal set of features with Support Vector Machine (SVM) as a classifier for classification and also K-mean which is used for the noisy data for classification. The k-mean procedure follows a simple and easy way to classify a given data through clustering [2][3] and is called as K-mean clustering algorithms. Genetic algorithms are used for feature selection which represents a feasible solution to a particular problem [4]. SVM is classier which performs classification tasks by constructing hyperplanes in a multidimensional space separating the case of different class levels [5]. Datasets have been used from Pima Indians Diabetes from UCI repository and the proposed model attained an average accuracy of 98.79 % for the reduced dataset.

**CNN:-**

*S. Lekha* has discussed and analyzed, A non-invasive method for detecting diabetes. In this paper, using an array of metal oxide semiconductor (MOS) gas sensors breath signals are measured required for data analysis and further, the data is analyzed using various pattern recognization technique. Pattern recognization technique which includes feature extraction and classification has been used to quantitatively analyze the VOC concentration and enhance the sensitivity of the sensor [2].In this study, the experimental data has been collected from the sensory unit. To classify signals obtained from an array of MOS sensors, a one-dimensional convolution neural network has been implemented. The **CNN** algorithm produces a set of feature maps by convoluting the raw data signals with a kernel-based filter. These feature extraction and classification systems are successful in diabetes recognition applications.

Logistic Regression :-

For early diagnosis and prediction of diabetes, *Changsheng Zhuet.al.* have proposed a data mining based model that is using the Pima Indian diabetes dataset. Main aim here is to determine ways of improving the k-means clustering and also logistic regression accuracy results to predict diabetes at an early. This model contains K-means, Logistic Regression and PCA (principal component analysis).To extract relevant information from a confusing data set, PCA is a simple, non-parametric method[2]. The result experimentally shows that the K-means clustering algorithm and logistic regression classifier accuracy is comparatively high than the published studies. One of the important issues solved is the accuracy of the prediction model. Using logistic regression here is an advantage as it can model a new dataset successfully.

The SMOTE (Synthetic Minority Oversampling Technique) given by *Manal Alghamdi et.al. ,* strategy is a sort of oversampling technique that has been demonstrated to be ground-breaking and is generally utilized in ML with imbalance high-dimensional information that is progressively utilized in medicine. The study shows that the **Logistic regression** classifier achieves the highest performance while the decision tree with the lowest performance. This study shows the capability of machine learning for predicting incident diabetes using cardiorespiratory fitness data. By exploring new machine learning models to increase the predicting quality more work could be done.

A data-driven cluster analysis was done by *Emma Ahlqvist et.al.( 2018)*  in patients from Swedish with newly diagnosed diabetes. To compare time to medication, risk of diabetic complications and time to reaching the treatment goal, and genetic associations, Cox regression, and logistic regression were used. Five identifiable replicable clusters of patients were studied which had the risk of diabetic complications and significantly different patient characteristics. The final result came was that the new clustering of patients with the classic diabetes classification was lesser than adult-onset diabetes. This study provides more precise data towards clinically useful stratification, of medicine in diabetes.

The results of multiple logistic regression analysis on the prediction of the odds ratio (OR) of cardiovascular risk factors indicate that an increased risk of diabetes is associated with increasing age, BMI, and blood pressure, and lower odds of developing diabetes were associated with male gender and having insurance coverage. According to the results of study by *Hossein Ebrahimi et.al* , increasing age increases the mean blood glucose and the prevalence of diabetes also increased accordingly. This multivariate logistic regression revealed a significant OR and the difference was statistically significant. It is highly recommended for paying attention to caring and controlling for diabetes in the third decade of life onwards Because of the increasing trend of diabetes incidence and the growing frequency of aging in Iran.

*Shankaracharya et. al . concluded* In the twenty-first century, diabetes has been recognized as a health challenge in developed as well as developing countries. Due to modernization, urbanization and economic development, diabetes frequency has been increased[1]. At present, the primary goal is to develop a diabetes diagnosis system using conceptual intelligence. Based on an artificial network and machine learning many approaches have been tested on the diabetes dataset[2]. The basic aim is to sketch out the range of development and potential machine learning algorithms as diabetes diagnosis tools. Application to diabetes diagnosis remains a challenge for conceptual intelligence after such rapid development. And this is due specific problem of data use which arises when statistical models of data are unknown or time-dependent, only partial data is available or when the parameters of the learning system need to be updated from time to time. It is suggested that within specific geographic regions models of the decision must be prepared on a dataset that intently represents the patient’s profile.

Date 10 se 2019

SVM

Kemal Polat and Salih Gunes, *et.al* proposed A cascade learning system for classification of diabetes disease using Generalized Discriminant Analysis and Least Square Support Vector Machine.The result obtained when using LS-SVM is 78.21% classification accuracy using 10-fold cross validation, the GDA–LS-SVM obtained 82.05% classification accuracy using 10-fold cross validation. The highest accuracy obtained was 82.05% and is very promising with regards to other classification. The fallout strongly suggest that Generalized Discriminant Analysis and Least Square Support Vector Machine classifier are based on a learning method can assist in the diagnosis of Diabetes disease.

T. Santhanam and M.S Padmavathi propose a work of Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis.Datasets have been used from Pima Indians Diabetes from UCI repository and the proposed model attained an average accuracy of 98.79 % for the reduced dataset.

Seokho Kang and Pilsung Kang proposed method utilizes data selection methods to lessen the training set before constructing a collection.The model for the prescription level calculation derived by our projected method achieved about 80% classification accuracy for prescription cases.

Muhammad Waqar Aslam and  Zhechen Zhu, *et.al* proposed a model using genetic programing technique with ultimate aim to facilitate the diagnosis of diabetes automatically by figuring out if a patient has diabetes, without the need of a physician.The Pima Indian diabetes dataset from the UCI Repository of machine learning databases has been used.The result shows that GP not only improves the act or the performance but also reduces the eight input dimensions to a single dimension.

Nahla H. Barakat, and  Andrew P. Bradley, *et.al* proposed a model of Intelligible Support Vector Machines for diagnosis, prognosis, and management of diabetes many of the data mining and machine learning methods.Results on the real-life prediction of type2 diabetes dataset show a comprehensible rule set which means intelligible SVM provides a promising tool for the prediction of diabetes with prediction accuracy of 94%, sensitivity of 93%, and a specificity of 94%

Bob Zhang and B.V.K. Vijaya Kumar, *et.al* proposed detecting diabetes mellitus and non-proliferative diabetic retinopathy using tongue color, texture and geometry features.By testing through each feature individually to healthy DM, the highest average accuracy gained was 66.26% via SVM. While employing SFS with SVM, the optimal result was shown by 9 features with an average accuracy of 80.52 %. when testing for NPDR DM the best result came was 5 features with 80.33% average accuracy.

……………………………………………………………………………………………………………….

IEEE

S. Lekha proposed a model where in the experimental data has been collected from the sensory unit by Real-Time Non- Invasive Detection and Classification of Diabetes using Modified Convolution Neural Network.The CNN algorithm produces a set of feature maps by convoluting the raw data signals with a kernel-based filter. These feature extraction and classification systems are successful in diabetes recognition applications.

Konstantia Zarkogianni and Maria Athanasiou , et.al proposed a model.The models were tested using the dataset from the medical record of 560 T2DM patients and the best discrimination performance noted was up to 71.48% in terms of AUC. the results obtained indicate that a hybrid ensemble integrating both the HWNN- and SOM- which are the primary model performs well even if the small number of CVD incidents were included.

………………………………………………………………………………………………………………………………………………………………….

Logistic Regression

Imran Kurt et. Al. has proposed  the classical arithmetic study between the presence and absence of CAD to examine the difference in the distribution of age variable with the basic motive was to Compare the performance of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease.

Emma Ahlqvist and Petter Storm , et.al proposed a model To compare time to medication, risk of diabetic complications and time to reaching the treatment goal, and genetic associations. A data driven analysis was done where in Cox regression technique, and logistic regression technique were used.The final result came was that the new clustering of patients with the classic diabetes classification was less than adult-onset diabetes.

The SMOTE(Synthetic Minority Oversampling Technique) was given by Manal Alghamdi and Mouaz Al-Mallah,et.al for predicting diabetes mellitus using machine learning approach This study shows that the Logistic regression classifier achieves the highest performance while the decision tree with the lowest performance.This study shows the capability of machine learning for predicting incident diabetes using cardiorespiratory fitness data

Changsheng Zhu and Christian Uwa Idemudia, et.al proposed a data mining based model that is using the Pima Indian diabetes dataset for early diagnosis and prediction of diabetes using k-means clustering and also logistic regression accuracy results.The result experimentally shows that the K-means clustering algorithm and logistic regression classifier accuracy is comparatively high than the published studies.

•Hossein Ebrahimi  & Mohammad Hassan Emamian, et.al proposed a model to determine the prevalence of diabetes and its risk factors among the middle-aged population.. According to the results of this study, increasing age increases the mean blood glucose and the prevalence of diabetes also increased accordingly.This multivariate logistic regression revealed a significant OR and the difference was statistically significant

•Mahmoud Heydari and Mehdi Teimouri, et.al aimed a model to compare the performance of several different classification techniques on a set of data that has been screened for type 2 diabetes. Based on the results obtained, it can be concluded that the efficiency of a model depends on the nature and complexity of datasets used.The most accurate method of classification in this research was an artificial neutral method with 97.18% accuracy.

•K. Saravananathan1 and T. Velmurugan analysed classification like algorithms J48, Support Vector Machines (SVM), Classification and Regression Tree CART and k-Nearest Neighbor (kNN)to best classify the input data of Diabetic dataset.The results show that the performance of the J48 technique is considerably better for the other three techniques for the classification of diabetes dataset.

•Shankaracharya and Devang Odedra,et.al reviewed artificial network and machine learning approch towards diabetes dataset with the basic aim is to sketch out the range of development and potential machine learning algorithms as diabetes diagnosis tools. and . It was suggested that within specific geographic regions models of the decision must be prepared on a dataset that intently represents the patient’s profile.

…………………………………………………………………………………………………………………………………………………………………