1. Introduction

Diabetes mellitus (DM) depicts a heterogeneous group of conditions that share certain features with hyperglycemia as a distinctive component. Also, lack of insulin secretion, insulin activity or both leads to hyperglycemia. If hyperglycemia of DM continues for a long period it introduces long-term damage to organs, their dysfunction, and consequent failures known as micro and macro-vascular complications. It leads to various serious problems such as heart attacks, eye defects, nerve damage and complication in the functioning of vital organs. Diabetes is increased blood sugar content in our body which is usually classified into two types i.e., Type-1 and Type-2 [1]-[3]. Type-1 diabetes is associated with an increase in hyperglycemia and majorly arises due to the deficiency insulin level [1]. Similarly, Type-2 diabetes will happen when the body cannot effectively use the insulin produced [2]. Usually, Type 2 diabetes is increasing at a fast pace due to obesity, physical inactivity, and unhealthy dietary habits [3] [2] [4]. The frequency of diabetes over the years has been increasing dramatically with the aging population worldwide. It increases the mortality rate and reduction in the life expectancy of elderly diabetic patients [2].

In the present day, various statistics confirm that diabetes is more prevalent in the urban areas i.e., 28%, while only 5% is observed in rural areas because of a sophisticated lifestyle and lot of junk food [ ]. As a result, diabetes patient has risen from 108 million in 1980 to 422 million in 2014. [2]. It is found in all age groups and killed almost 1.6 million people in the year 2016. This figure exhibits that DM is replaced by HIV/AIDS and becomes the seventh top reason for death in early-stage [1]. Over the period, it becomes a pandemic disease and affects nearly 415 million people around the world [ ]. Recent studies show that 80% of type 2 diabetes complications can be prevented through timely intervention [2], [4], for example, by changing their lifestyle [3] or by therapeutic methods. A large number of diabetic deaths can be checked, if it can be diagnosed as much as earlier. Hence automated classification tasks are playing an important role in predicting and diagnosing diabetes. It will assist a physician or dietician in diagnosis or narrow-down the diagnosis and medication accordingly. It will not only improve the diagnosis accuracy but also helps patient time and wealth. In such a situation patient has to provide blood samples during the fast as well as after diet which is painful and tedious because a patient has to visit various labs which are followed by doctors. Sometimes, anxiety and workload of the physician also affect the identification of disease or trauma level and prescribing usual dosages of medicine which hampers a lot to the patient health. 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.

In fact, 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. At present, the primary goal is to develop a diabetes diagnosis system using conceptual intelligence. This intelligent system is required to consider the various factors and identify a suitable model between the different parameters so that computational complexity, programming complications as well as accuracy of diabetes classification can be enhanced. Based on an artificial network and machine learning many approaches have been tested on the diabetes dataset [2]. However, application to diabetes diagnosis remains a challenge for conceptual intelligence after such rapid development. In fact, various machine learning classification techniques (such as K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Kernel SVM, Naive Bayes, Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), Random Forest (RF)) have been developed to classify the diabetes [ ]-[ ].

The main contributions of this paper are outlined as follows.

1. The diabetes data are collected from Modern Diabetes Hospital, Ranchi, India for the experimentation purpose.
2. The proposed classification parameters can be learned offline, and after achieving the optimum value, it can be applied for different online diabetes dataset.

The layout of the paper is organized as follows: Section 2 describes the literature review of diabetes classification using machine learning techniques. Proposed methodology is discussed in Section 3. Similarly, outcomes of different experimental task are presented in Section 4 and eventually result discussion and conclusion of this research article are exhibit in Section 5 and Section 6 respectively.

**2. Related Work**

The upcoming most popular and important approach in the field of medical research is Machine Learning. In clinical diagnosis problems, classification plays a vital role in further treatment of the disease. 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. In this section is to sketch out the series of development and outline a comparison study of potential machine learning algorithms as diabetes classification tools. 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.Various studies have been performed on the diabetes data classification. In this literature review section we are going to discuss the work of different researchers done on all the classification of disease using intelligence system, which are as follows:

* 1. **K-Nearest Neighbor (KNN)**

Krati Saxena *et.al.,* developed a model for classification of Diabetes Mellitus using K-Nearest Neighbor algorithm and calculated accuracy and error rates for K=3, 5. The result is shows that as the value of k increases, accuracy rate and error rate will also increase. Likewise, M.Akhil *et.al* formed a model wherein classification of heart disease using K-Nearest Neighbor and Genetetic algorithm. The performance of this model is tested with 6 medical data sets and 1 nonmedical dataset. Out of 7 datasets, 6 datasets were chosen from UCI Repository [ ]. Result shows that accuracy of heart disease data decreases 32% using cross validation method whereas Accuracy of the heart disease is increased by 5% using GA full training data set and 15% improvement in accuracy for cross validation against KNN without GA.

Y. Angeline christobel and P. Sivaprakasam has worked on a model for a new class wise K-Nearest Neighbor (CKNN) method for the classification of diabetes dataset. The dataset has been taken from PIMA Indian diabetes from the UCI machine learning repository .The data set has 768 instances with two class problems to test whether the patient having diabetic or not. The performance of classification is measured with respect to specificity, sensitivity and accuracy has been increased notably in the case of CKNN algorithm.

Similarly, using PIMA, Nirmala Devi. M *et.al.,* has suggested a model an amalgam KNN to predict Diabetes Mellitus. Model has been tested with medical dataset downloaded from UCI machine learning data repository for diabetes mellitus that is PIDD. The results showed that the performance of K-Means-KNN model can be further upgraded by multi-step pre-processing using deleting the missing values and grouping by K-means clustering with KNN classification.

C. Kalaiselvi and G .M. Nasira modelled a new approach for classification using PIMA data for training and testing the neural network model for prediction of diabetes and cancer using ANFIS. Many researches were conducted to predict both diabetes and cancer. The experimental results shows that the classification accuracy is better than existing approach taken. The ANFIS approach gives reduces complexity and higher efficiency.

**2.2 Logistic Regression :-**

For early diagnosis and prediction of diabetes, Changsheng Zhu *et.al.* have worked on a data mining based model that is using the PIMA. 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) reported 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.

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 discussion of results 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.

Imran Kurt *et. al*. have investigated  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 *et.al* developed 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.

Changsheng Zhu *et.al,* worked on a data mining based model that is using the PIMA 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 *et.al*, examined 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 *et.al* investigated 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. Saravananathan 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 *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.

**2.3** **Support Vector Machines**

*Seokho Kang* (2015) examined 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.

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.  SVM’s 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 discussed 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., have discussed the classificationof diabetes disease and so a new cascade learning system is been investigated 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 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.

*T. Santhanam , M.S Padmavathi* reported that 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.

Kemal Polat *et.al.,* designed 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 experimentally investigated a work of application using PIMA of K-means and Genetic Algorithms for dimension reduction by integrating SVM for Diabetes diagnosis. The investigated model attained an average accuracy of 98.79 % for the reduced dataset.

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

Muhammad Waqar Aslam *et.al.,* have worked upon 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 PIDD 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. Baraka *et.al.,* modeled 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 *et.al.,* contributed in 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.

**Naïve Base**

Yue Huang a *et.al.,* examined feature selection and classification model on type 2 diabetic patients. Methodology include naïve Bayes [12], decision tree learner [13] and instance-based learner[14] and feature selection result applied in real world, accurately classified models.

Rashedur M. Rahman and Farhana Afroz conducted a comparison of various classification techniques using different data mining tools for diabetes diagnosis. Methodology used includes Multilayer Perceptron, Bayes Net [12] ,Naïve Byes, J48graft, Fuzzy Lattice Reasoning, Adaptive Neuro-Fuzzy Inference System and Performance Metrics. Naïve Bayes  classifier has the highest accuracy above all.

Aiswarya Iyer *et.al.,* reported classifications of diabetes using mining techniques like wise using decision tree and Naïve Bayes algorithm. The results show that 70:30 percent split for Naïve Bayes gives least error and the model prepared is quite effective.

Dewan Md. Farid and Li Zhang *et.al.,* investigated diabetees using Decision tree, Naïve Bayes as methodology. The NB classifier obtained average accuracy rate of 86.7%. Naïve Bayes Tree (NBTree), rough set approach and fuzzy logic are accurate and will be implemented in future works .

**Random Forest**

Beatriz López *et.al.,* developed methodology for Artificial intelligence in medicine And compared RF vs SVM and LR. The complexity, overfitting, interactions were handled by RF and it is a useful method for predictive models when compared to SVM and LR.

Guoyan Zhaoa *et.al.,* conducted a study for intestinal virome changes precede autoimmunity in type-1 diabetes-susceptible children Random forest analysis is used to identify disease-associated virus.

Attila A. Seyhan *et.al.,* presented the Pancreas-enriched miRNAs altered in the circulation of subjects. random forest classification is used to estimate the diagnostic odds ratio (DOR) and sensitivity analysis is implemented using the R environment.results: The prediabetic group showed a significant reduction in miR-126 and miR-146a. T2D showed elevated levels of miR-30d, miR-21, miR-34a. T1D exhibited high levels of miR-21 and miR-375.

**Decision Tree**

Gaganjot Kaur and Amit Chhabra, investigated a modified J48 classifier to increase the accuracy rate of the data mining procedure. The data mining tool WEKA has been used as an API of MATLAB for generating the J-48 Classifier. It clearly states that the porposed algorithm has large accuracy difference than other algorithms. It has accuracy rate of 99.87% rather than others that show maximum  has been proposed of 77.21% accuracy.

Asma A. AlJarullah , a method is reported to predict the patients with developing diabetes. The dataset used in this is PIDD, which collects the information of the patients with or without developing diabetes. WEKA's J48 Decision Tree classifier was applied to the modified dataset to construct the decision tree model. The accuracy of the resulting model was 78.1768%

**FLANN**

Humar Kahramanli and Novruz Allahverdi developed a model for the classification with the existing records using the simple data. The classification acquires of this dataset were obtained by k-fold cross-validation. The proposed method achieved accuracy values of 84.24% and 86.8% for the PIDD and Cleveland heart disease dataset, respectively.

Satchidananda Dehuri *et. al.,* proposed an improved swarm optimized functional link artificial neural network (ISO-FLANN) for classification and author evaluated algorithm using a set of fifteen public domain datasets from the University of California at Irvine (UCI) machine learning repository. Experimental study shows that the performance of ISO-FLANN for classification task is promising. Results obtained with the EFLSNN model proved to be as good as well as better than the best results found by the MLP, SVM, FLANN with gradient descent and FSN.

CNN:-

S. Lekha has discussed and analyzed, A non-invasive method for detecting diabetes. In this study, 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].Investigation of 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.

S. Lekha developed 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 *et.al.,* examined 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.

1. Methodology
2. Proposed Methodology.
3. Results
4. Discussion

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

1. Conclusion

References