1. **What is Diabetes, types of Diabetes and its complications on human Body?**
2. Introduction

Diabetes mellitus (DM) depicts a heterogeneous group of conditions that share certain features with hyperglycemia as a typical component. Lack of insulin secretion, insulin activity or both leads to hyperglycemia. If hyperglycemia of DM continues for a long time it introduces long-term damage to organs, their dysfunction, and consequent failures known as micro and macrovascular complications and leads to various serious problems such as, heart attacks, eye defects and nerve damage. Diabetes is increased blood sugar content in our body and it is usually classified into two types i.e., Type1 and Type2 [1]-[3]. \*Add 2 line for type1 and type2\*. (First Para) Diabetes is associated with an increase in hyperglycaemia [1]. One reason for this increase is insulin deficiency. This is known as type 1 diabetes. The type 2 diabetes is where the body cannot effectively use the insulin produced [2]. 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. Increased mortality rate and reduction in the life expectancy of elderly diabetic patients are due to growing diabetes [2].

(Second Para) Various statistics show that diabetes is more prevalent in the urban areas is 28%, while only 5% in rural areas [ ]. In fact diabetes patient has risen from 108 million in 1980 to 422 million in 2014. [2]. At present, it is found in all age group and it killed 1.6 million people in the year 2016. This figure exhibits that DM is replaced HIV/AIDS and becomes as the seventh top cause of death [1]. It become a pandemic disease and affects around 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. Large number of diabetic death can be checked, if it can be diagnose as much as earlier. Hence classification is an important task in predicting and diagnosing the diabetes.

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]. \*\*\* Why ML? 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. Machine learning classification methods such as Decision Tree (DT), Naïve Bayes (NB), Logistic Regression (LR), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), Kernel SVM, Naive Bayes, Random Forest (RF) Classification.

\*\* Your Contribution & Paper Structure.

The basic aim is to sketch out the range of development and comparison study of potential machine learning algorithms as diabetes diagnosis tools.

1. 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].

Literature Review

1. Methodology
2. Regression
3. Neural Network
4. SVM..................
5. Proposed Methodology.
6. Results
7. Discussion
8. Conclusion

References

The upcoming most popular and important approach in the field of medical research is Machine Learning[1].

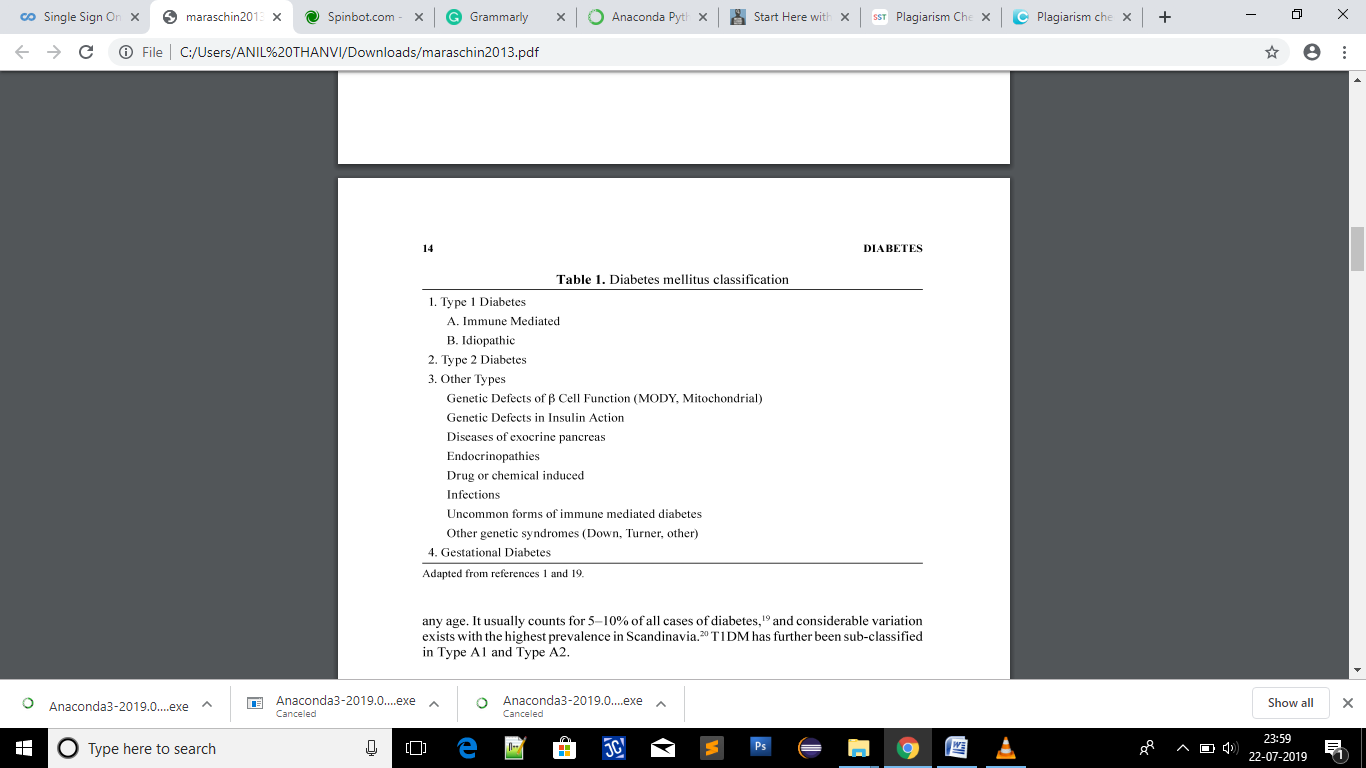
Diabetes results in abnormal increase of glucose level in the blood, which damages the functioning of organs mainly kidney, heart and brain.[4] Effects of diabetes can be suppressed when it is identified in the earlier stages.

. [5] Diabetes mellitus (DM) is a group of metabolic disorders posing significant threat to human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data.

**CLASSIFICATION OF DIABETES**

There is a vast growth of DM in pandemic areas. Further study shows that there will be countries like India, China, and the United States with the most number of people diabetic by the year 2025 -2030. There is a very rear case of reduced incidences of microvascular complication but the correct diagnosis can of DM control the level of glycated hemoglobin in both Type 1 and Type 2 diabetes. The American Diabetes Association (ADA) and the American Heart Association (AHA) recommends Hemoglobin below 7%.[4][5][8][9][11][13]

**CLASSIFICATION OF DIABETES MELLITUS**

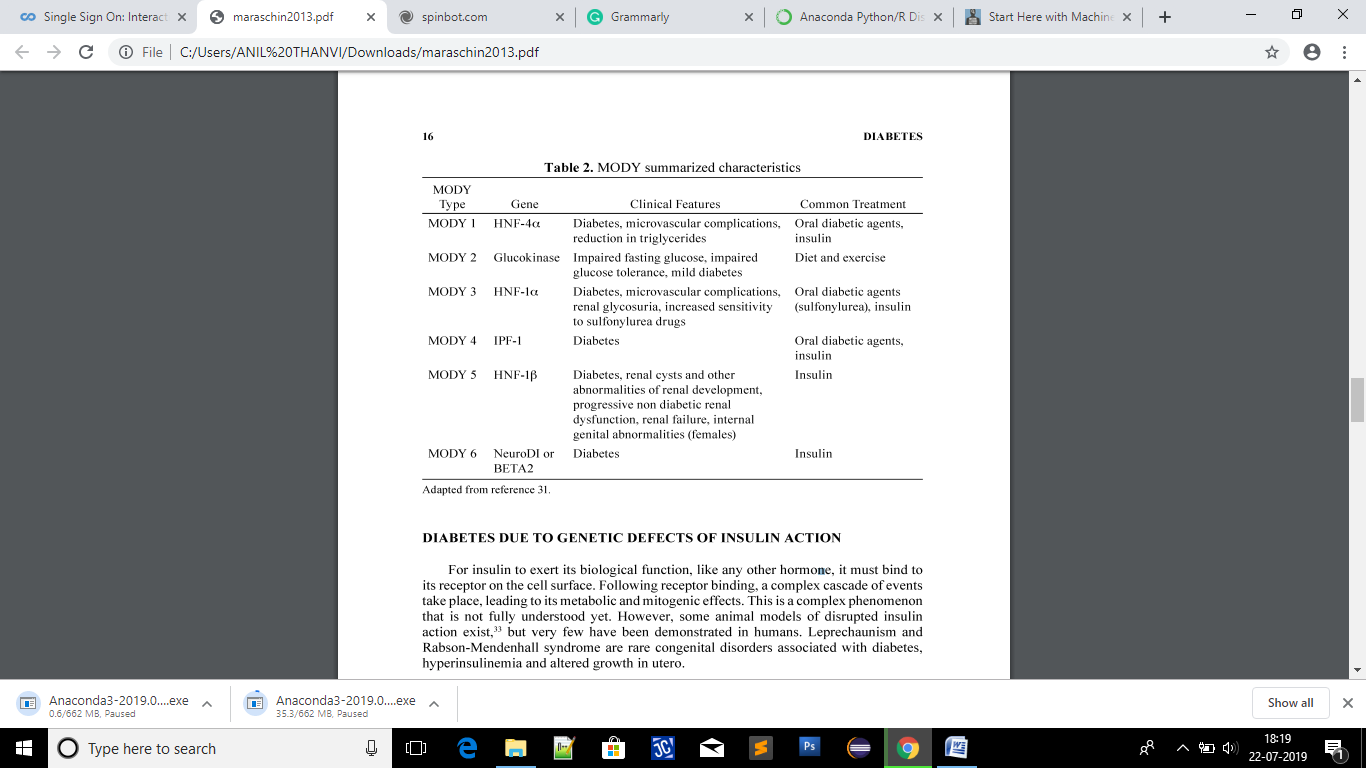
In 1979 by the National Diabetes Data Group (NDDG) the first classification of diabetes was published and was endorsed by WHO. In this Insulin-dependent diabetes mellitus (IDDM) and non-insulin-dependent diabetes mellitus (NIDDM) groups were derived based on pharmacologic therapy. But by the period some problems became evident, which lead to misclassifying the patients. The correct control and treatment of the disease can only be achieved by the correct classification. The new classification was endorsed by the WHO and modified by ADA based on the pathogenesis of the disease and not its treatment and there Four major categories were planned, Type I diabetes mellitus, Type 2 diabetes mellitus, other specific types of diabetes and gestational diabetes.[15][18][16][17]

**TYPE 1 DIABETES:** It is a chronic disease in which complete insulin deficiency. The patient dies out if deprived by insulin. This disease can occur at any age but common in childhood and adolescence.[19][20]

* **AUTOIMMUNE OR TYPE 1A DIABETES**: This is an organ-explicit immune system issue with a hereditary inclination dictated by the significant histocompatibility locus on chromosome 6[21]-[24]
* **IDIOPATHIC OR TYPE IB DIABETES**: lt is normally depicted in African or Asian ethnic gathering; it needs HLA affiliation and are prone to ketosis. They were classified as "ketosis-prone diabetes”[25]

**TYPE 2 DIABETES**: An important feature of this disease is Progressive Beta-cell failure which is the most common type of diabetes which corresponds to 90% of the worldwide.[19][27][28][29][30]

**DIABETES OF OTHERS SPECIFIC TYPES**: Many types exist but the major one is maturity-onset diabetes of the young (MODY) which has been better defined genetically.

* **MODY(maturity-onset diabetes of the young) :** There are six major types of MODY, each with its atypical phenotype and hereditary characteristics.[31][32]
* **Diabetes Due To hereditary Defects Of Insulin Action:** Leprechaunism and Rabson-Mendenhall disorder is an uncommon intrinsic issue related to diabetes, hyperinsulinemia and altered development in utero. Binding of insulin to its receptor on the cell surface is mandatory to perform its biological function, like any hormone.[33]
  + *Type A insulin resistance: It is a syndrome of insulin resistance in females*
  + *Type B insulin resistance* : It is a rear type of syndrome[34]
* **Diabetes Due To Exocrine Pancreas**
* **Diabetes Due To Other Endocrinopathies :** Insulin action can be disaffected by Cortisols, the growth hormone, glucagon, and catecholamine. The interesting part about this is that once tumors are removed, diabetes can disappear.
* **Drug- Or Chemical-Induced Diabetes:** Beta-cells can be affected by certain chemicals and can even be toxic and is irreversible.[37]
* **Diabetes Due To Infections :** Viruses like Rubella and coxsackie B can cause diabetes[38][39]
* **Gestational Diabetes Mellitus (GDM)** : This type of diabetes occurs at the beginning or during pregnancy. The risk is greater with older age, hereditary, body mass index before pregnancy. its acknowledgment is essential because under treatment can prompt opposed results to the mother and fetal. [40][41]

1. ***Existing Methods for finding Diabetes***

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.

1. **Why MACHINE LEARNING Techniques used today:-**

Machine Learning is making machines learn from experiences .Some scientists believe that machine learning is equivalent to Artificial Intelligence as learning is the main characteristic of any intelligent machine. The main aim of ml is to form systems that can learn through experiences [3]. According to Mitchel [4],"A computer program is said to learn from experience E with respect to some tasks T , measured by P ,improves by gaining experience E." Importance of ml is growing in the field of medical sciences through various machine learning methods like: Logistic Regression , K-Nearest Neighbours (K-NN) ,Support vector machines(SVM),Kernel SVM Random Forests , Naive Bayes , Decision tree to predict diabetes in the early stages.

The SMOTE (Synthetic Minority Oversampling Technique) 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. ( logistic regression –Anirudhi)

**4). What are techniques used for diabetes classification?**

In terms of machine learning, classification is an instance of supervised learning.

A Classification problem is a problem in which we separate our input data into discrete categories. By discrete, we mean separate classes.

Unlike regression where we predict a continuous number, we use classification to predict a category. There is a wide variety of classification applications from medicine to marketing.

Category is a group of people or things similar in some way.

For examples - \* Whether a mail is spam or not. Which category it belongs to?

\* To assign a diagnosis to a given patient, based on observed characteristics of a patient.

Classification models include linear models like Logistic Regression, SVM and nonlinear ones like K-NN, Kernel SVM and Random Forests.

In this part, we will understand and learn how to implement the following Machine Learning Classification models:

1. Logistic Regression
2. K-Nearest Neighbors (K-NN)
3. Support Vector Machine (SVM)
4. Kernel SVM
5. Naive Bayes
6. Decision Tree Classification
7. Random Forest Classification

**Literature Review:**

( logistic regression –Anirudhi)

Diabetes is the fastest growing disease universally and a substantial threat to human health [1]. There are two types of Diabetes, Type1 diabetes, and Type 2 diabetes, but type 2 diabetes, in particular, is highly assorted[2]. A data-driven cluster analysis was done 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.

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

People suffering from diseases do not realize at an early stage about their health conditions and disease like diabetes causes a large number of deaths each year [1]. For early diagnosis and prediction of diabetes, we have proposed a data mining based model that is using the Pima Indian diabetes dataset. Our 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.

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

Diabetes mellitus is a major challenging health problem around the world[1]. This study intended to determine the prevalence of diabetes and its risk factors among the middle-aged population in Shahroud, north of Iran[2]. Simple and multivariate logistic regression were used to determine the risk factor associated with 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 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. 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.

…………………………………………………………………………………………………

One of the worldwide increasing diseases is diabetes and the disease will subsequently get increasing if countries do not consider preventive measures for disease[1][2]. 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 type 2 diabetes, in the city of Tabriz, Iran. The compared classification techniques were a support vector machine, artificial neural network, decision tree, nearest neighbors, and Bayesian network. Based on the results obtained, it can be concluded that the efficiency of a model depends on the nature and complexity of datasets used. To obtain the best possible result the choice and selection of features for data mining to be done by the help and advice of the expert . The most accurate method of classification in this research was an artificial neutral method with 97.18% accuracy.

…………………………………………………………………………………………………

In our study, To predict the presence of coronary artery disease, performances of classification techniques were compared.  Coronary artery disease (CAD) is the most important global health issue with its incidence and death rates. The significant hazardous factors for CAD are hereditary, smoking, hypertension, hypercholesterolemia, diabetes mellitus, and obesity. One of the most interesting and difficult tasks is to develop data mining applications by predicting the outcome of the disease. We have formed models using LR(Logistic Regression), CART(Classification And Regression Tree), neural network algorithms which are often used for classification problems. LR is helpful for circumstances in which you need to have the option to predict the presence or absence of a character based on the factors of independent variables which are consistent, categorical, or both. CART can deal with numerical information that is highly skewed or multi-modal. Neural networks have been utilized to model medical and functional outcomes of the hazardous malady. We performed the classical arithmetic study between the presence and absence of CAD to examine the difference in the distribution of age variable. By the analysis of the Shapiro–Wilk test  Age was tested for normal distribution. And the Mann–Whitney U-test performed the Comparison between two groups for non-normally distributed age variable. For presence and absence of CAD nominal variables were tested by the chi-square test.

( logistic regression –Anirudhi)

**………………………………………………………………………..**

# (K-Nearest Neighbor Algorithm- Raghav)

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

One of the major worldwide health issues is Diabetes, around 346 million are suffering from Diabetes Mellitus(DM) which is characterized by Hyperglycemia, globally according to the report by WHO in 2011. The conversion of sugar into simple molecules is affected by the poor functioning of the Pancreas, which results in the accumulation in the bloodstreams, therefore the blood sugar level is increased in the body. It is a chronic disease that may lead to severe damage to the nerves and blood vessels. Common symptoms are excessive thirst, hunger, sudden weight loss or weight gain, etc. this can also lead to heart attacks and visual impairments. KNN algorithm uses the nearest data points for the estimation; therefore it is easy to take complete advantage of the local data and create highly nonlinear, highly adaptive boundaries of decision for every data point,[1] used cascading K mean ad KNN algorithm for classification of the patients, [3] have introduced a much faster K-Nearest Neighbor algorithm for text categorization. With this, K-Nearest Neighbor can be used for the diagnosis of Diabetes Mellitus.

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

Missing data is the main problem for the data quality, imputation of the mean of the data is the only solution for the missing values. In this paper, a new method for the classification of Diabetes Dataset is being used i.e, Classwise K Nearest Neighbor (CKNN) method. The negative impact of the missing values in the data was later improved by the improvement in the KNN algorithm. KNN algorithm has effectively been used in various applications of data analysis[5,6] for e.g, retrieval of the information, database, recognition of the pattern. According to the Pima India Diabetes Dataset, 32.4% is the error rate for the KNN method and the average miss classification rate is 26.20%. In [12], Jeatrakul and Wong have done a comparative study on five different types of neural networks. 24.05% is the average rate. Twala Jones and Hand[13] created a method for preparing a separate class for the missing data. Satheesh et al [15] designed a dynamic nearest neighbor classifier for integrated data through object-oriented concept and generalization. Data preprocessing is very important for providing quality results, where [20] the imputation of the missing values in Machine Learning is challenging. To impute the missing values mean substitution was done, to increase the accuracy of KNN and also propose a CKNN. Mean substitution, Data Normalization,  K-Nearest Neighbor(KNN) Classification Algorithm, The proposed CKNN Algorithm, Validating the Performance of the Classification Algorithm were the algorithms which were discussed. Implementing and evaluating the proposed CKNN classification algorithm, and measured the average performance of the algorithm.

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

The amalgam model combines K-Means with KNN(K-Nearest Neighbor) with multi-sep preprocessing. By removing the noisy data, the accuracy and efficiency of the KNN algorithm and the quality of the data are improved. The larger the K is, the lesser will be the effect of noise on the classification. The aim is to determine the value of for PIDD for better classification accuracy using an algorithm. According to the report of the WHO, a person is dying every 10 seconds in the world. The major percentage of Diabetic patients in the world will be present in the developing country of the world like India. India is the Diabetes capital of the world, around 35 million Diabetics. Data Mining is done to predict various diseases based on the previous data collected from the patients. The Interpretations made by the Physician are essential components of these data[7]. The data present in the real world tends to be incomplete, inconsistent and noisy, to fill the missing data, Data Preprocessing is needed[6[8][9]. The values are unknown so the correction of the data is difficult, though we are using the values of mean, clustering algorithm, range, or standard deviation. If the value of K is too large, it may result in increasing the computational complexity[11]. The results are further compared by statistical computing measures such as accuracy, specificity and sensitivity are used as the three criteria in the medical field. With this, amalgam KNN is proposed as one of the best results compared to the results which were reported in the literature.

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

A common metabolic disorder is Diabetes Mellitus (DM). Peripheral vascular disorder and Diabetic neuropathy are closely related to Diabetic Foot Ulcers (DFU's). Amputations and Lower Extremity Ulcers are ongoing crises among individuals with diabetes. With an increasing number of patients and the prevalence of 6.4% globally, DM is the most common metabolic disease. The most serious, alarming, and expensive complications and compromising the well being and the survival of the diabetic patient(3). Collection of the demographic data of patients including the details about their history of diabetes and the other factors like BMI, Blood Pressure, etc. Complete healing means ulcer closure and no dressing is required(11). Amputation is of 2 types, Minor Amputation, and Major Amputation, where Major Amputation is also known as Transtibial Amputation(19). A retrospective study was proposed via chart and photographic review, with keeping a record of the patients who visited the hospital from January 2010 to December 2014. Univariable Logistic Regression analysis was conducted to examine the association between the classification with amputation and the candidate predictors. Backward selection for the initial multivariable logistic regression model was determined by this method. The substantial accuracy of the system and their main variable are associated with LEA occurrence.      This work was supported by the Soonchunhyang University Research Fund.

**(K-Nearest Neighbor Algorithm- Raghav)**

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

(Slot-1 Bhanvi-Random forest)

**Diabetes is a kind of metabolic disorder characterized by hyperglycemia and abrupt proportions of carbohydrates that is the source of high sugar contents in our body.the 2 types of diabetes insulin-dependent(type-1)and non-insulin-dependent (type-2) both require the application of data mining techniques to a database for effective management of diabetes.**

**Intelligent data analysis techniques such as feature ranking and classification model construction are used to extract previously unknown but useful information for decision making and predictions.classification methods used in the research paper are probabilistic learners, naïve Bayes [12], decision tree learner [13] and instance-based learner[14]. feature selection is used to remove the tons of unwanted and meaningless information so, feature ranking is used to rank top features required for diabetes prediction. Classification techniques determine the total number of accurately classified models. the results suggest that the model suggestive predictions could be applied to routine practices and widely acceptable for medicinal practitioners.**

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

**In this paper, we study handling large data sets through clustering and classification techniques which can be resolved by machine learning techniques that involve the study of data mining tools such as WEKA, TANAGRA, and MATLAB. The dataset is taken from the UCI repository[9] of prima Indians diabetes data for  diagnosis of diabetes in prima Indians .methodology used are ,Multilayer Perceptron [11] which is a classification algorithm used for neural networks , Bayes Net [12] ,Naïve Byes[12] (for probabilistic prediction of data) ,J48graft, Fuzzy Lattice Reasoning(Flr),(for decision making rules in ml), Grip(J Ripper),Fuzzy Interface System(Fis)(Deals With Fuzzy Logic),Adaptive Neuro-Fuzzy Inference System and Performance Metrics(used for measurement purposes). The results from the above methodologies were formulated and it showed that the Naïve Bayes  classifier has the highest accuracy followed by WEKA J48graft then ANFIS IN MATLAB.**

**………………………………………………………………………………………………………………………………**

## Diabetes is a deadly disease in today’s world and has engrossed 246 million people already, the majority of them being women. In this paper, we mainly focus on the impact of diabetes on pregnant women through the approach of decision tree and Naïve Bayes algorithm on our data set to classify and predict diabetes. Diabetes is of 3 types type-1, type-2 and gestational diabetes [2]with type-2 being the most common of all diabetes. Early identification is essential for curing any type of diabetes. Some common symptoms of diabetes are frequent urge to urinate, repetitive hunger and thirst tendencies and strange weight loss. 10-fold cross-validation is used to prepare test and training set for model. The results show that 70:30 percent split for Naïve Bayes gives least error and the model prepared is quite effective. Further improvements in the research work can make it suitable for concluding diabetes.

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

**In this paper, hybrid algorithms are used for decision tree and naïve bayes classifiers and recent research for real-world applications of DT and NB. Data mining algorithms are required to construct classifiers. Classification is a kind of data mining operation used to predict class labels of instances while clustering is a grouping of things that have a high degree of resemblance .Decision tree is used to classify instances in large datasets. Naïve Bayes is a simple and effective classification method in data mining and ML. The DT algorithm finds out the troublesome training instances and then the NB classifier removes those instances from the training set. Hence the accuracy of the dataset is maintained. In the second model , the most important subset of attributes is found through DT and their weights are calculated . Evaluation of performance of the proposed algorithm is done against the existing traditional models using parameters like accuracy, precision, sensitivity, and specificity analysis , and 10 fold cross-validation on datasets taken from UCI ML repository .Results show that the proposed DT models outshine the traditional ones and  increases accuracy rate by an average of 4.8% .The NB classifier obtained average accuracy rate of 86.7% on 10 datasets using 10fold cross-validation. In conclusion algorithms such as naïve bayes tree (NBTree), rough set approach and fuzzy logic are accurate and will be implemented in future works .**

(Slot-1 Bhanvi-Random forest)

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

(Slot-2 IEEE Journals –Anirudhi)

Support Vector Machines

Diabetes mellitus is a never-ending disease and is merely associated with an increase in the level of blood glucose [1]. Here a novel hybrid model for medical diagnosis has been proposed, integrating three different data mining and machine learning techniques. In this study for diagnosis, prognosis, and management of diabetes many of the data mining and machine learning methods are used.  SVMs work by finding a linear hyperplane that separates the positive and negative examples with an utmost interclass distance. SVM cannot give intelligible support for the classification decisions they make, which means they are black-box models and so we used the intelligible representation of the SVM’s diagnostic. SVM to be a more intelligible rule extraction technique has been introduced. Dataset used can be found in [2] which previously investigated the prevalence of diabetes in 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].

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

Slot-2 –Bhanvi Springer Journal

The Cost Of Type 2 Diabetes Mellitus : A Machine Learning Perspective

In this paper burden of diabetes is quantified by calculating the indicating features that discriminate between normal people and diabetic patients. The type II diabetes mellitus (T2DM) is engrossing over 312 million people who need to be hospitalized due to several complications. The major problem is that the health care data is widespread, noisy, complex and missing at certain points hence making it difficult for us to analyze by conventional machine learning algorithms. The effective cost of T2DM on patients' life is determined through ml techniques. The dataset used is derived from the Practice Fusion De-Identified Data Set which provided information of 10,000 patients [11]. The various feature sets applied are Baseline Features, Diagnosis Features, Medication Features, and Lab Test Features. The evaluation of these features is based on their discriminative power. Classification is done through Logistic Regression and Support Vector Machines. The quantification of feature-specific contribution increases performance measurement. While the Lasso Ridge regression [16] is not useful to describe the main features. It is concluded that the burden of diabetes is not limited to a small test, it is beyond these features and needs deep investigations.

Diabetes prediction in health care systems using machine learning algorithms on hadoop cluster

The rapid development in the field of Information and Communication Technologies is responsible for remarkable improvements in health care monitoring system. various machine learning algorithms are deduced to automate models and increase their accuracy. The Hadoop cluster is designed for storing and comprehending huge amounts of unstructured data [7]. It is a complex cluster that divides the task to various cluster nodes that work parallel to analyze the big data[8]. Big data was put for the first time in 2009 and has multiple uses in the medical field[8]. Due to large and voluminous data available cluster-based computing is more effective. To develop the diabetes prediction model data of 75,664 patients is used to take from the National Institute of Diabetes .70%of the data is used as a training set and rest 30/5 as a test dataset. Information gain is used as a feature selection model that improves the accuracy of the model by reducing noise ( Irrevelant features that influence classifier ). Neural Networks learn from the examples during training based on the existing examples. SVM attempts to finding a hyperplane separating the classes they are also known as large margin classifiers and are quite complex. The training dataset is built into a decision tree while the naïve Bayes classifier makes the data train faster. Random forests correct overfitting. the overall accuracy of different machine learning algorithms produces 6% accuracy with random forests producing the highest accuracy. The Hadoop cluster performs much better when compared to the other two machine learning algorithms.

Classification of diabetes

Diabetes mellitus (DM) is currently posing an epidemic problem all over the world. The number of affected patients is expected to rise to 300 million by 2025. India, China, and the United States are estimated to find the largest diabetics by 2030. The National Diabetes Data Group(NDDG) classifies DM into 2 major groups: insulin-dependent diabetes mellitus(IDDM) and non-insulin dependent diabetes mellitus(NIDDM). And also into (a) gestational diabetes, (b) malnutrition-related diabetes. The TYPE-1 diabetes is due to complete insulin deficiency patients may even enter a coma and die if left insulin deprived. It is most commonly found in children or teenagers. The T1DM is further subdivided into type A1 and A2. The organ-specific autoimmune disorder involving genetic predisposition is determined by histocompatibility locus on chromosome 6[21]. The idiopathic or 1B diabetes is due to the lack of autoimmune basis to beta cell destruction[25]. Type 2 diabetes is most common in 90% of the cases and is found in adults after 40 years of age due to obesity. various other types of diabetes include Maturity-Onset Diabetes of the Young(MODY) it involves regulation of insulin secretion .there are various other diabetes due to insulin action, exocrine pancreas, endocrinopathies, drug-induced, infections and gestational.

Diagnosis of diabetes type-2 using hybrid machine learning based ensemble model

In this manuscript, an intelligent framework for doctors is proposed for correctly diagnoses Diabetes mellitus. The expert system developed known as “Diabetes Diagnoser” (this even works on small mobile [5])is powered by 4 AI algorithms that have been tested through 10 fold cross-validation and results compared with actual patients. 15 classifiers are considered out of which 5 major ones are: ANN, SVM, KNN, Naïve Bayes and Ensemble. The ensemble is an enhancing technique that combines the classifying ability of individual classifiers and chances of miss classifying is decreased.GUI based diagnosis tool would predict if the patient is suffering from a disease when it is fed with all 10 deciding attributes through GUI which is developed with the help of MATLAB. A multi-layer feed-forward neural network (MLFFN) and genetic algorithm (GA) is used to predict heart disease. ANN is trained using backpropagation methods and feed-forward neural networks showed an accuracy of 79.7% on a training set and 89.67% on test set[7,8]. Dataset is obtained from the UCI ml data repository. The methodology includes 2 steps, first Feature selection algorithms reduce data size from 8 to 4 most important attributes. These are then fed to the neuro-fuzzy interference system for AI-based diagnosis. The artificial immune recognition system-2 (AIRS2) [12] had a version of AIRS2 called MAIRS2 which made use of fuzzy KNN which help improve diagnosis with 89.10% accuracy than 82.69 in AIRS2. The overall efficiency of the diagnostic tool is 97.34% which can be increased by increasing number instances

Accurate diabetes risk stratification using machine learning: role of missing value and outliers

Diabetes mellitus is the cause of 1.5 deaths directly and 2.2 deaths indirectly due to cardiovascular diseases in 2012[3]. In this paper, we objectify to develop a robust and efficient machine learning system under the assumption that all missing values are replaced and hence will yield high-risk satisfaction[10-16]. Optimization of data set is done by selecting the best features and classifying them based on the set of 6 feature selection techniques and 10 classification models. Conventional techniques used are LDA (linear discriminant analysis), QDA(quadratic discriminant analysis), NB(naïve Bayes), GPC(Gaussian process classification), SVM(support vector machine), ANN(artificial neural network), Adaboost, LR(logistic regression) and DT(decision tree). The parameters for analyzing the are ACC, SE, SP, NPV, PPV, and AUC. The database is taken from the UCI repository containing 768 females with a variety of specifications. In the methodology preprocessing of raw data takes place by considering missing data. Next is Feature selection which improves the quality of the dataset and reduces computational cost and training time of the model. The various FS process PCA(principal component analysis), analysis of variance, Fisher discrimination ratio, logistic regression, and random forests. Using the above listed 10 classification features experiments were performed but random forests(RF) showed the most significant features and accurate diabetes prediction it works well on both non-linear and high dimensional data. RF produces very meaningful results that correlate risk factors too and yields 89% accuracy.

Classification of diabetic patient data using machine learning techniques

Health care services require the best predictions and persistence in the models as these can be life threatening. Data mining techniques are used to derive knowledge from large datasets and diagnose diseases. The training set contains observations that are associated with a particular category. The classification techniques used are Association rule learning and Artificial Neural Networks.ANN is implemented through the Waikato Environment for Knowledge Analysis(WEKA). The output is completely based on the accuracy of the model. Data preprocessing is explicitly done using python as it has various helpful libraries such as pandas: for data analytics and matplotlib: for graphical representation of test and training data sets on the model. Also, the Jupyter notebook (IPython3) is used for interactive performance. The Association Rule Learning (ARL) based on discovering relations between tuples in large databases [7]. R is used to implement statistics into the model through the Arles library. dataset derived from UCI which is de-identified concepts from[8] consists of 10520 train set and 4509 test sets. similar and equally accurate results are shown by both the methods ARL and NN because both are data mining techniques and are efficient.

A survey of medical diagnoses of diabetes using machine learning techniques

In this paper, the leading lifestyle ailment characterized by elevated blood glucose levels is diagnosed using various algorithms. Diabetes mellitus is of the following types: Type-1, Type-2, pregnancy diabetes, diabetic retinopathy, and diabetic neuropathy. Major factors responsible could be overweight, genetics, high mercury contents, environmental factors, low level of HDL[1]. Successful machine learning algorithms immensely help to solve various health associated issues and determine diseases at an early stage which can increase speed, performance, reliability, and accuracy of diagnoses for the disease. Techniques used to classify patients at either “less risk” or “high risk” stage. SVM[4] method is applied to deduce a hyperplane that partitions the data into different classes and removes the overfitting of the sample. k-NN is lazy learning wherein “k” nearest neighbors [5] are identified and the anonymous data point is classified according to neighbors. DT [6] a graph-like structure containing a target variable is formed and traverses from root to leaf for taking the decision. RF [7] is a collection of few DT and it is a dimension reduction method for easy analyzing of multi-dimensional problems. naïve Bayes [8] is quite simple and used for probabilistic classification. Artificial neural network [9] consists of input, hidden and output layers that are also called as multilayer perceptron(MLP). Logistic regression(LR) model [10] is a statistical approach to ml models. All the above-mentioned algorithms are performed on the PIMA Indian Diabetic dataset[18] on various parameters and gave in to the result that LR is most accurate for the classification of diabetic and nondiabetic samples.