International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-2, December 2019

An Ontology Driven System to Predict Diabetes With Machine Learning Techniques

Divakar H R, D Ramesh, B R Prakash

Abstract: ***Diabetes Mellitus is considered one of the chronic diseases of humankind which causes an increase in blood sugar. Many complications are reported if DM remains untreated and unidentified. Identification of this disease requires a lot of physical and mental trauma and effort which involves visiting a doctor, blood and urine test at the diagnostic center which consumes more time. Difficulties can be over crossed using the trending technology of Machine learning. The idea of the model is to prognosticate the occurrence of a diabetic with high accuracy. Therefore, two machine learning classification algorithms namely Fine Decision Tree and Support Vector Machine are used in this experiment to detect diabetes at an early stage. Therefore two machine learning classification algorithms namely Fine Decision Tree and Support Vector Machine are used in this experiment to detect diabetes at an early stage.***

Keywords **:** ***Diabetes Mellitus, Ontology, Fine Decision Tree, SVM, Machine learning.***

I. INTRODUCTION

In the recent decades of the current global world, there are so many chronic diseases present in medical field. Every rise of the day there is a presence of new diseases and its complications are reported, both in the developing and developed country such serious disease is widely spread throughout the world. One among those serious diseases reported in the humankind is Diabetes Mellitus [DM] widely known as "sugar" in India. It is evident from recent statistics and research that India is majorly affected by a dangerous and most common disease which has affected children and adults is diabetes.

Majorly DM is classified as two classes medically recognized as T1DM and T2DM respectively. Type 1 DM (T1DM) occurs when your immune system i.e when body's resistance fails which inversely attack and destroy the insulin-producing beta cell of the pancreas. Scientists assume type1 diabetes is caused by genes and ecological factors, such as virus, that may elicit the disease. Pancreas produces insulin the produced insulin is regulated by hormones in the body which in turn maintains the sugar level in the body, when this balance is affected, body is affected by the disease. Type 2 DM (T2DM), the most common type of ailment that occurs due to lack of physical activity, due to this blood glucose level became imbalance, this is recognized as blood sugar. Presence of complex glucose in blood is the main source of energy and comes mainly from food a human consumed. The insulin, one of the important hormone produced by beta cells of pancreas, helps to absorb glucose and converts as energy. In this, body reacts in two types, body fails to produce sufficient insulin or body fails to utilize the insulin produced. Too much glucose is retained in your blood in the second type or not enough insulin is reached to your cells in the first case. It is the most common one caused when body cells stop responding to insulin's produced in the body. Type – 2 diabetes has significantly increase in the current time, according to International Diabetic Federation Projection of prevalence of diabetic is expected to reach 471 million by 2035, that means one of ten people will be suffering from diabetic[1,2].

Research has proved that diabetic is a root cause for many other diseases like, Diabetic Retinopathy, Macro Vascular Complications, Nephropathy, Neuropathy. It is very much important to identify the diabetes in the early stages. Diabetic has become very common in children and adults in the current lifestyle and food habit. The early identification can help in understanding the disease completely by which educating in improving sedimentary life style, food style which can be used to maintain diabetes. Currently blood sugar level and urine sugar level are measured to detect the diabetes. These parameters to detect diabetes have failed as symptoms of diabetic are ambiguous and the procedure involved in detecting is painful, concerning children [3].

There are numerous ways to predict diabetes in early stage in the growing technological world. The emerging technologies used in prediction are Artificial Intelligence, Machine Learning, Deep Learning along with the help of ontology. As diabetic is dominating there is a huge quantity of data that produce from medical history of diabetes patients, there is massive interest in extract useful information and discover the hidden patterns. An information tool provides the ontology elucidation in healthcare domain [4]. The advantage of large amount of data that generate by such diseases, reclaim the others acquaintance to classify current needs in order to improve the upshot in the prospect iterations [5]. It needs a formal version of diabetes jargon, terms and interaction to ascertain, mine, carve up, retrieve
and reuse knowledge. In this ontology [6] is one of the techniques to accumulate semantic in a row and facilitate manipulate with data by applying diverse methods such as analysis and algorithms [7].

In the midst of the advance of livelihood standards, diabetes is progressively more common in day by day life of humans.

Revised Manuscript Received on December 30, 2019.
***** Correspondence Author
Divakar H R*, PES College of Engineering, Mandya, Karnataka, India. (Email: divakarhr@gmail.com)
Dr. D Ramesh, Professor, Sri Siddartha Academy of Higher Education, Tumkur, Karnataka,India. (Email: rameshd_ssit@yahoo.com)
Dr. B R Prakash, Govt. First Grade College, Tiptur, Karnataka, India. (Email: brp.tmk@gmail.com)

Consequently, how to hurriedly and perfectly identify and explore diabetes is a subject matter worth study. Various algorithms are used to foresee diabetes together with the established methods. Machine learning method is most far and wide worn in predict diabetes and they get preferable accurate consequences. We propose an ontology driven model to predict the occurrence of the T2DM in human beings.

II. LITERATURE SURVEY

Comparable increasing in the quantity of individuals, WHO affliction from polygenic disorder, there's escalating within the associated complications as a development of polygenic disorder. This paper associate's metaphysics as primarily based model, considering adult patients those who are affected and plagued by polygenic disorder. To find the foundations of polygenic disorder with its complications in disease and disease relationship the data technology tool metaphysics is employed as an answer supplier which contributes care towards chronic disease patients affected by diabetes. Here, they developed associate metaphysics for adult patients plagued by polygenic disorder in Saudi Arabia to foretell patient that have middling or menace altitude with sure impediment. It facilitates care supplier to go looking and gift them with history concerning patient facts like force per unit area, diagnosis report, and science laboratory check, varied convention between diverse jeopardy factors of polygenic disorder of patients, surely generate complications. By using these data new complications may be revealed as new discovery of the study, discover polygenic disorder complication which may be helpful to postponement of complication [4].

The treatment of T2DM is a multifaceted problem. A CDS system based on the considerable and disseminated e-record of health data can ease the mechanization of this procedure and augment the exactness. The most imperative constituent of any CDS system is its acquaintance. This acquaintance can be formulated by means of ontology's. The prescribed depiction logic of ontology ropes the presumption of concealed acquaintance. Building a absolute, articulate, unswerving, interoperable and sharable ontology is a confront. Here, they introduce a new ontology for predict the treatment for the DM patients. Diabetes Mellitus Treatment Ontology (DMTO) based on the diabetic domain, its standard and interoperable knowledge relevant to treatment of T2DM. It adheres to the blueprint principles suggested by the open Biomedical Ontologies Foundry and is based on ontological pragmatism that follow the principles of the Basic Formal Ontology and the Ontology for broad-spectrum therapeutic science. This ontology is competent to accumulate and analyze most characteristics of T2DM as well as adapt unceasing treatment procedure with the most suitable drugs, food and substantial exercises [1,12].

In humans DM could be a major reason for morbidity and mortality, early findings are the beginning toward the management of polygenic disorder. An identification of this involves many variables, that makes it tough to gain correct and also timely identification and construction of accurate unceasing treatment procedure. An e-health documentation system needs an integrated call uphold aptitude, and ontology that now turning into needed for economical, consistent, extendibility, reusability and semantically intellectual data. They develop a theory which is sound and semantically intelligent having knowledge of domain for resolution issues associated with the identification of polygenic disorder. Such data can modify and act as replacement of categorial patient central clinical call support systems which will facilitate attention supplier to detect diabetics rapidly and correctly. DM identification metaphysics provide a typical ontology which will sustain ability among clinical call support system and attention systems. It's in an exceedingly one amongst in every sickness metaphysics residential to signify numerous ailment aspect in typical coherent form [2].

Decision Tree plays a major role in prediction by using classification and regression methodology. They are used in classification of medical data and use it for prediction. The idea of decision is based on if-then rule, using a tree structure beginning with a single node, then node getting expanded as a leaf [3,8]. Random forest is a combination of a greater number of Decision tree and it is a multi-functional algorithm, after training the samples unseen samples can be found from the trained samples [4]. Neural network has two types feed-forward and back-propagated neural network. It has many layers naming input layer, output layer and hidden layer which has an activation functions like sigmoid, remu and many more. Using these methods, the data was collected and a model is built, the model is validated using hold-out and k-fold method. The data set is divided to k sections, the sections are named as folds. K-1 sections are used to train the model and one section is used to test the model [8]. After validation feature selection method PCA and mRMR is used to lessen the attributes selected. Sensitivity, Specificity and accuracy and Mattews correlation coefficient is used to measure the accuracy. Using these accuracy methods against the data set it was found that collecting all the features and using MRMR resulted in higher accuracy than PCA results. During this it was proven Machine Learning can predict the diabetic but choosing the correct model, method and validation is important.

The physiological monitor of adult citizens by means of wearable sensors has shown great impending in humanizing their eminence of life and preventing undesired measures related to their healthiness. Hence, the building vigorous foretelling model from data composed inconspicuously in habitat can be a challenge, principally for elderly aged inhabitants.

To perform this task, they offered a method to recognize basic substantial activities from wearable sensors, with reverence to challenges arise from gadget generated or human allied parameters. Data from elder participants with dissimilar levels of defenselessness and purposeful situation be used to guide and review the end to end model outline developed to deal with those challenge. Categorization was performing by typical machine learning, as well as deep learning technique, indicating a slight improvement of the last. On the whole domino effect sustains the use of the anticipated motion recognition system for inconspicuous monitor of elder citizens [9].

Kadam Vinay et.al., [10] used deep learning to predict the disease on patient treatment history. Previous models were able to predict the main stream diseases and not the sub set diseases and also, they could handle only the structed data not the unstructured data sets.

They proposed a system which used Stochastic gradient descendants and Artificial neural network [ANN] to predict the disease in both structured and unstructured data. ANN acts as a human barin in prediction having multilayer network. The data set contained the treatment history of diseases like kidney, heart and diabetes. They combined all the data set to a common data set and along with the data they combined other factoring attributes affecting these diseases, the missing values and were generated using linear regression and decision tree. ANN is built for the data set using keras libraries of python, which uses dense and sequential model to build. The weights were allotted randomly and the result obtained was given as input to the next layer and so on till the final layer. The result is compared with the actual result, errors are calculated. The procedure is carried out in back-propagation with updated weights. Support Vector Machine is applied on the data set by diving it into training and testing frames. The predictions are made accordingly.

Irene Dankwa-Mullan, MD, MPH,Marc Rivo, et.al., [11], made a research on various Artificial Intelligent technology used in reforming the diabetic care. According to the study made the clinical data is hazardously unstructured from investigation to discharge history. There is huge of amount of data which is getting generated daily in clinical field including screening of images, laboratory reports, patient records, non-clinical data etc. It is difficult to handle them effectively, but according to the various research 450 papers analysed using Artificial Intelligent technique could manage this unstructured, ambiguous data effectively. The study classified the diabetic care into four major areas such as, ARS, CDS, foretelling inhabitants risk stratification and serene self-management tool. It has been concluded that AI is prominently showing the best result and in return the implementation of AI in the field has increased. It is because AI has the ability to interpret and manage the huge number of data efficiently, it has high efficiency in managing screened images, reports. AI has large number of algorithms which helps in learning the relationship among the data. Cognitive AI is much more efficient in understanding as it goes deeper layers and predict the output in every layer. Deep learning of AI has given the 90% accuracy of predicting harm of eyes using retinal photography.

III. METHODOLOGY

a. Ontology driven machine learning model

The process of ontology development is one among concept to determine the associations and properties in the realm of diabetes mellitus. The intention of this ontology generation is to foresee the occurrence of the diabetes in females of PIMA Indians data base. Protégé 4.3 is used to construct the DM ontology for the Pima Indians diabetes dataset via top-down procedure. The FaCT++ and HermiT reasoners are used to test out its reliability.

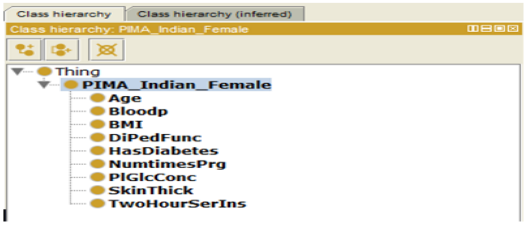


Figure 1: Class hierarchy of Pima Indian diabetes ontology.

Figure 1 depicts class hierarchy of the ontology. This ontology is built for dataset initially gathered from National Institute of Diabetes and Digestive and Kidney Diseases. The intention of this study is to foretell based on the investigative dimensions whether a patient is suffering from DM or not. Based on several constraints the dataset is created by selection of various instances. In this dataset, all patients are females and age group greater than 21 years of Pima Indian heritage. Each instance of these records contains information about how many times they were Pregnant, Plasma glucose concentration, Diastolic blood pressure, Triceps skin fold thickness, Two-hour serum insulin, Body mass index, Diabetes pedigree function, Age and the criterion is on discovering patient instance is diabetic or healthy based on the available data.

The main object properties that mapped class and subclasses are has age, hasbloodp, hasbmi, hasDipedfunc, hasNumtimesPrg, has PlglcConc, hasskinthick, hasTwohourserins. The diagnose of diabetes is determined by 1 and absent by 0. The data property determin the classes value in the ontology where various criteria are identified.

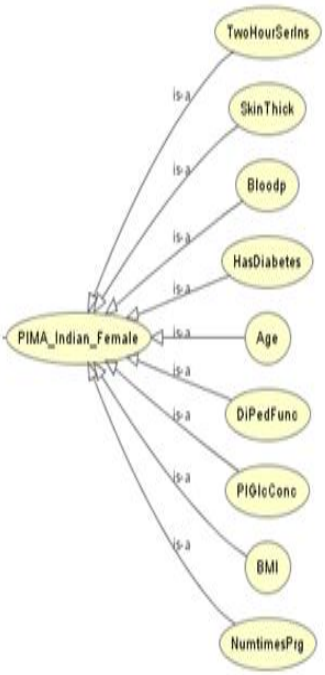


Figure 2: Asserted class hierarchy and the inferred class hierarchy produced by Protègè

Figure 2 demonstrates the class pecking guidelines in an OWL cosmology to be seen and incrementally explored, permitting inspection of the asserted class sequence of control and gather class evolution. Appraisal phase is one it improve the formed beliefs after various instance of time and essential of the end user. Documentation step is successively perform in equivalent with all distinct other phase where each of the movements happen to construct the ontolgy and each one of its perspectives are kept back. Figure 3 depicts the tree formation by OWLViz module in Protègè.

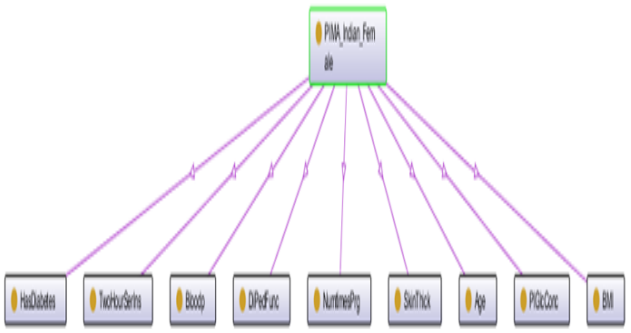


Figure 3: DomainOWL tree structure produced by Protégé

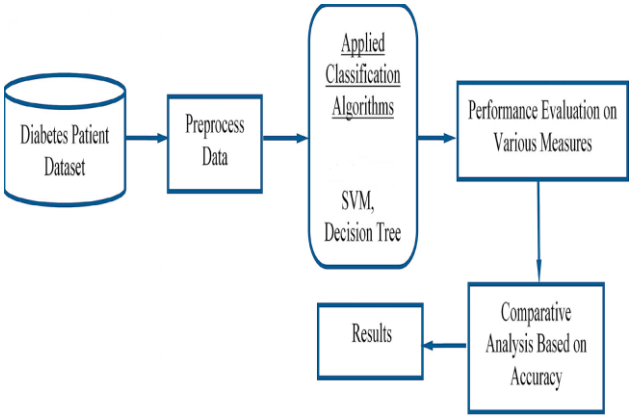


Figure 4: proposed prediction model diagram

The figure 4 summarizes the procedures used in the proposed work as a model. This shows the flow of research study conducted and depicts the flow of work undergone.

B. Classification Algorithms

Fine Decision Tree algorithm

A DT is one among the supervised machine learning algorithms [8]. Where it builds a tree structure where each division node represent a alternative between a number of alternatives, and each sibling node represents a decision. Information theory is used to determine to delineate the grade of inadequacy in a system known as Entropy.

Entropy is a measure of superfluous data from a source of messages. Given a group S, contains positive and negative examples of a few aim notion, the entropy of S is relative to this categorization.

$$\text{Entropy(S)} = \sum_{i=1}^{c} - p_i \log_2 p_i \text{----------1)}$$

Where, p_i is the proportion of S belonging to class i

Information gain decides which feature goes interested in a decision node. To reduce the decision tree depth, the characteristic with the most entropy diminution is the best option.

$$Gain\ (S,A) = Entropy\ (S) - \sum_{v\ \in\ Values\ (A)} \frac{[S_v]}{[S]} Entropy\ (S_v) \in \text{--- (2)}$$

Where: S is each value v of all possible values of attribute A

Sv = subset of S for which attribute A has value v

|Sv| = number of elements in Sv

|S| = number of elements in S

Support vector machine

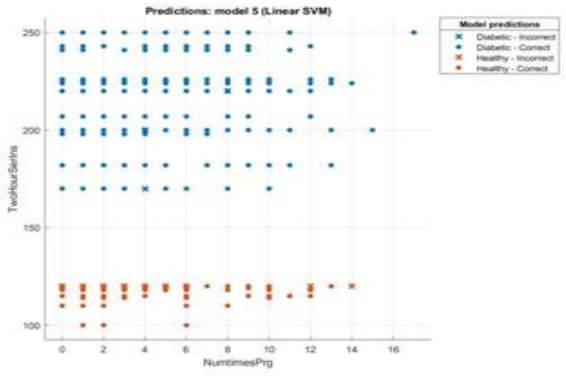
SVM is one of the supervised learning algorithms in machine learning area. It is used to analyze the records which is used for classification and regression analysis. This algorithm effectively performs the non-linear classification and also map the inputs in to the high dimensional feature space which is used for classification, detection and regression.

Dataset

Here we utilize the Pima Indians diabetic's dataset.The patients considered here are females of age greater than 21 years elder of Pima Indian tradition. This dataset contain about 8 attributes which are identified at the time of pregnancy, attributes are like plasma glucose level, bp, triceps skin fold thickness, two hour serum insulin level, bmi, DM prediction function and the age of the female patient. The original data set contains 786 diabetic data insistences.

IV. VALIDATION OF PROPOSED MODEL USING MACHINE LEARNING ALGORITHMS

The data was non-inheritable exploitation the PIDD keep into the memory. The recorded knowledge was pre-processed and options were extracted from the stable region of the detector transient response. The chosen options were normalized and then fed to the input of the fine call tree and SVM using machine learning formula. The trained network models 1 and 2 are developed in the using fine decision tree and SVM. The predicted output is matched with the targeted output and fetched the results as '1' assigned for diabetic female patient and '0' which assigned for Healthy female from Pima dataset.



(a)

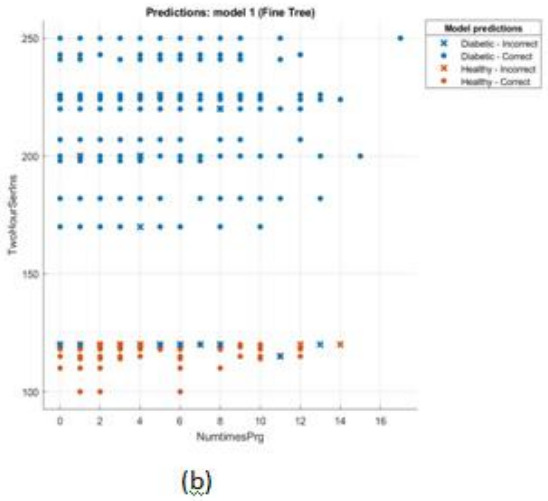
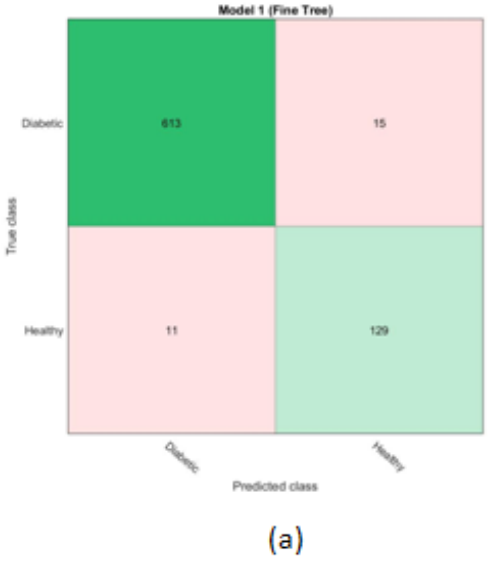


Figure 5: Scatter plots (a) fine decision tree algorithm (b) Support vector machine



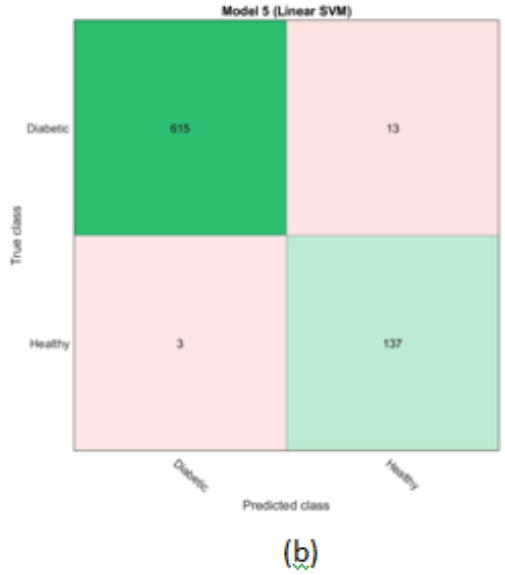
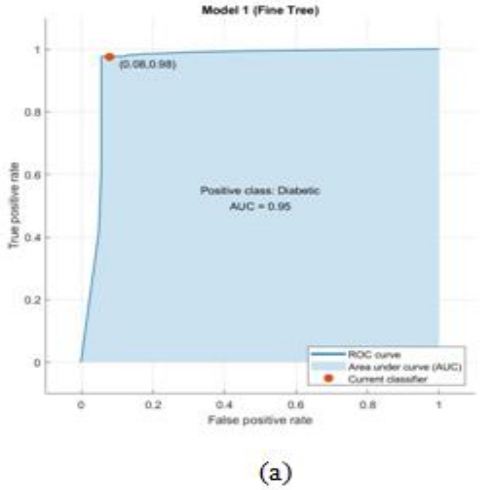


Figure 6: Confusion matrix of predicted output of (a) fine decision tree algorithm (b) Support vector machine



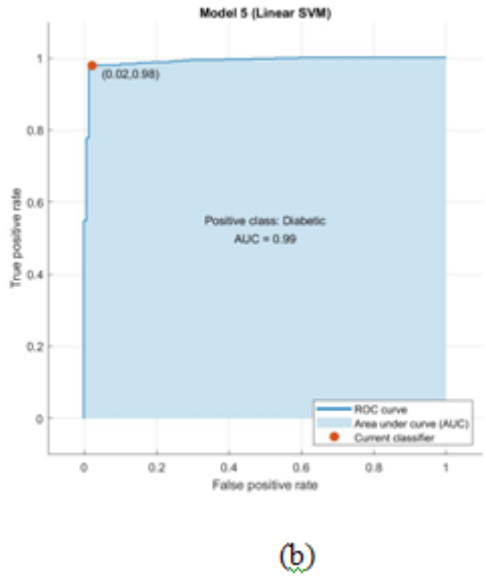


Figure 7: ROC of (a) Fine decision tree algorithm (b) Support vector machine

After that testing set is applied to the trained model of the algorithm, results obtained are plotted as shows in figure 5. Figure 5 showing scatter plot provides a visual representation of the correlation, or relationship between the two variables and it gives the clear separation of '1 correct' assigned for Diabetic samples is show in blue color,'2 correct' assigned healthy samples show in saffron color. Figure 6 shows confusion matrix of diabetic and healthy samples used to illustrate the performance of a classification model on a set of experiment data for which the right values are known and used to give the classification accuracy of the predictive model.

a. Measurements

In this crammer we used sensitivity (SN), Specificity (SP), Accuracy (ACC) and Matthews correlation coefficient (MCC) to measure the classified efficiency. Where PP is the total number of female patient samples having diabetes, NP is the total number of female patient samples those are non-diabetic, TP is true positive, FN is false negative,

FP is false positive and TN is true negative. To accomplish this following equations are used.

$$SN = \frac{TP}{TP+FN} \quad \text{---------- (3)}$$

$$SP = \frac{TN}{TN+FP} \quad \text{---------- (4)}$$

$$ACC = \frac{TN+TP}{TN+TP+FP+FN} \quad \text{---------- (5)}$$

$$MCC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}} \quad \text{---- (6)}$$

Table 1: Sensitivity and specificity of detection of diabetes

Classifier		Test positive	Test Negative	Total
FDT	Diabetic	613 (TP)	15 (FN)	628 (PP)
	Healthy	11(FP)	129(TN)	140 (NP)
SVM	Diabetic	615 (TP)	13 (FN)	628 (PP)
	Healthy	3(FP)	137(TN)	140 (NP)

Table 2 : Prediction of diabetes considering 8 features.

Classification Algorithm	SN	SP	ACC	AUC	MCC
Fine Decision Tree	0.9761	0.9214	0.9661	0.95	0.8878
Support Vector Machine	0.9792	0.9785	0.9791	0.99	0.9328

V. DISCUSSION

To discuss the overall performance of the proposed model. An ontology model was developed for the characteristic data set of Pima Indian origin female patient. Table1 provides the classification of instances of dataset used. For this experimentation a total of 768 patient records was considered. We build and compile the model by using machine leaning algorithms.

Table-2 determines classifiers accomplishment on the basis of categorization of available instances. According to these categorized instances, accurateness is determined and analyzed. Achievement of these algorithms is appraised based on the cor-rectly classified instances and imperfectly classified instances out of a entirety of instances. Figure 8 shows the effectiveness of these classification algorithms on the basis of classified instances. From Table-2 we can conclude that SVM classification algorithm outperforms compare to Fine Decision Tree algorithm. So, SVM is well thought-out as the best supervised machine learning technique of this experimentation because it gives superior accuracy in respective to other classification algorithms with an accuracy of 97.91 %.

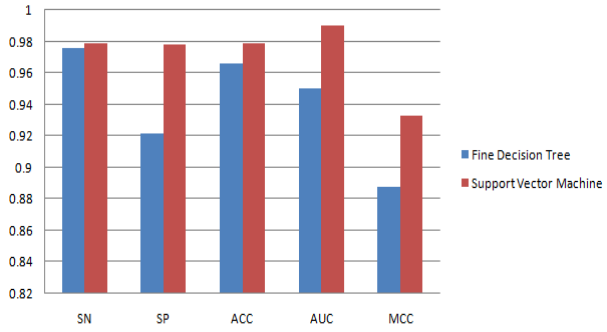


Figure 8: Various measures of classified effectiveness

VI. CONCLUSION

One among the various major problems in medical field is the recognition of diseases like DM in early state. In this cram, well thought-out efforts are made in designing a method which fallout in the forecast of diabetes. All the way throughout this work, an ontology driven machine learning classification algorithms are studied and evaluated on a range of measures. Experimentation are conducted on PIDD, experimental results determine the accuracy of 96.61% using Fine Decision Tree algorithm and 97.91% using Support Vector Machine, machine learning algorithms. In upcoming approaches, the designed method using machine learning classification algorithms can be improved by adopting deep learning concepts and can be used to improve prediction and identification of other ailment. The work can be unmitigated and enhanced for the automation of DM analysis together with some other machine learning and deep learning algorithms.

REFERENCES

1. El-Sappagh, Shaker & Kwak, Daehan & Ali, Farman & Kwak, Kyung. (2018). DMTO: A realistic ontology for standard diabetes mellitus treatment. Journal of Biomedical Semantics. 9. 10.1186/s13326-018-0176-y.
2. El-Sappagh, Shaker & Ali, Farman. (2016). DDO: a diabetes mellitus diagnosis ontology. Applied Informatics. 3. 10.1186/s40535-016-0021-2.
3. Czmil, Anna & Czmil, Sylwester & Mazur, Damian. (2019). A Method to Detect Type 1 Diabetes Based on Physical Activity Measurements Using a Mobile Device. Applied Sciences. 9. 2555. 10.3390/app9122555.
4. Daghistani, Tahani & Alshammari, Riyad & Razzak, Muhammad. (2015). Discovering Diabetes Complications: an Ontology Based Model. Acta Informatica Medica. 23. 385. 10.5455/aim.2015.23.385-392.
5. Boulos MK, Harvey FE, Roudsari AV, Bellazzi R, Jha MK, Pakhira D, Chakrborty B. Diabetes detection and care applying CBR techniques. International Journal of Soft Computing and Engineering (IJSCE). 2013; 2(6): 132-137.
6. Pramono D, Setiawan NY, Sarno R, Sidiq M. Physical activity recommendation for diabetic patients based on ontology. 7th International Conference on Information & Communication Technology and Systems. 2013; 27-32.
7. Divakar H R, B R Prakash & Mamatha M (2019). An Ontology Based System for Healthcare People to Prevent Cardiovascular Diseases. International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, Volume-8 Issue-2s11, pp 983-988.
8. Zou, Quan & Qu, Kaiyang & Luo, Yamei & Yin, Dehui & Ju, Ying & Tang, Hua. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics. 9. 10.3389/fgene.2018.00515.

9. Papagiannaki, Aimilia & Zacharaki, Evangelia & Kalouris, Gerasimos & Kalogiannis, Spyridon & Deltouzos, Konstantinos & Ellul, John & Megalooikonomou, Vasileios. (2019). Recognizing Physical Activity of Older People from Wearable Sensors and Inconsistent Data. Sensors. 19. 10.3390/s19040880.
10. Kadam Vinay R, K.L.Soujanya & Preety Singh. (2019). Disease Prediction by Using Deep Learning Based on Patient Treatment History. International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, Volume-7 Issue-6, pp 1159-1168.
11. Dankwa-Mullan, Irene & Rivo, Marc & Sepulveda, Marisol & Park, Yoonyoung & Snowdon, Jane & Rhee, Kyu. (2018). Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here. Population Health Management. 22. 10.1089/pop.2018.0129.
12. El-Sappagh, Shaker & Alonso, Jose & Ali, Farman & Ali, Amjad & Jang, Jun-Hyeog & Kwak, Kyung. (2018). An ontology-based interpretable fuzzy decision support system for diabetes diagnosis. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2852004.
13. Dorcely, Brenda & Katz, Karin & Jagannathan, Ram & Chiang, Stephanie & Oluwadare, Babajide & Goldberg, Ira & Bergman, Michael. (2017). Novel biomarkers for prediabetes, diabetes, and associated complications. Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy. Volume 10. 345-361. 10.2147/DMSO.S100074.
14. Chen, Rung-Ching & Jiang, Hui & Huang, Chung-Yi & Bau, Cho-Tsan. (2017). Clinical Decision Support System for Diabetes Based on Ontology Reasoning and TOPSIS Analysis. Journal of Healthcare Engineering. 2017. 1-14. 10.1155/2017/4307508.
15. Alehegn, Minyechil & Joshi, Rahul & Mulay, Preeti. (2018). Analysis and prediction of diabetes mellitus using machine learning algorithm. International Journal of Pure and Applied Mathematics. 118. 871-878.
16. Burns,G.A., Li,X. and Peng,N. Building deep learning models for evidence classification from the open access biomedical literature. Database (2019) Vol. 2019: articale ID baz034; doi:10.1093/database/baz034.