Machine Learning Techniques and Data Extraction Approaches in Diabetes Healthcare: A Comprehensive Review

Afshan Fatima[1], Saurabh Pal[2], Venkateswara Rao Ch[3]
[1]Research Scholar, CSE Department, Vbs Purvanchal University
[2]Professor, CSE Department, Vbs Purvanchal University
[3]Associate Professor, Department of CSE, Siddhartha Institute of Engineering& Technology
[1]afshafatima0889@gmail.com
[2]drsaurabhpalvbspu@gmail.com
[3]chvenkatsrh@gmail.com

Abstract: The contemporary world finds itself grappling with the pervasive impact of diseases, and diabetes stands at the forefront. As per the "International Diabetes Federation", a staggering 246 million individuals worldwide currently live with diabetes, and this figure is projected to soar to a monumental 380 million by the year 2025. This metabolic disorder, characterized by the mismanagement of blood glucose levels, engenders a heightened susceptibility to an array of ailments, including heart attacks, kidney disease, and renal failure. In light of these concerns, healthcare practitioners necessitate a dependable prognostic methodology to effectively diagnose diabetes mellitus. Fortunately, the rapid strides made in the realm of Machine Learning and Data Mining present a plethora of techniques and algorithms within the domain of artificial intelligence that can be harnessed with efficacy for disease prediction and diagnosis. This comprehensive paper endeavors to furnish a discerning review of the machine learning and data mining methods routinely employed in the analysis and prognostication of diabetes.

Keywords: Artificial Intelligence, Machine Learning, Data Mining, Diagnosis, Diabetes, Healthcare.

1. Introduction

Machine Learning and Data Mining are poised to revolutionize a multitude of fields, including the realm of healthcare [1]. As these tools continue to advance and become more prevalent, their potential to aid in the diagnosis of a wide range of diseases, including diabetes mellitus, becomes increasingly evident. The early detection of diabetes is a critical challenge, and data mining has emerged as a crucial player in diabetes research, harnessing the wealth of available data and uncovering hidden knowledge [2]. The application of various machine learning and data mining techniques has proven invaluable in diabetes research, leading to enhanced healthcare outcomes for patients and physicians alike.

In the pursuit of accurate treatment, a systematic approach to diagnosis is imperative in the realm of medical science. The rapid progress in computer technology, coupled with the advancements in machine learning and data mining, has inspired researchers to develop software solutions that assist doctors and patients in making informed decisions.

This research article delves into an extensive review of existing literature, exploring the diagnostic and predictive applications of data mining and machine learning in the context of diabetic healthcare. The primary objective of the authors is to construct a predictive model for diabetes mellitus, leveraging the power of machine learning and data mining methodologies.

The organization of this review is as follows: The Introduction, found in Section 1, sets the stage for the subsequent sections. Section 2 provides a comprehensive overview of the fundamental concepts underlying data mining, machine learning (ML), and deep learning. Section 4 delves into the different types of diabetes mellitus, while Section 5 presents an in-depth analysis of the publications reviewed in this study. Finally, Section 6 concludes the review, offering insights into future possibilities and areas of further exploration.

Data Extraction:

Data extraction, referred to as Knowledge Unearthing within Information Repositories (KUIR), serves the purpose of unearthing valuable insights from extensive databases or data repositories. The applications of data extraction span across both commercial and scientific domains [3].

Data extraction is defined as the intricate process of unearthing previously concealed patterns and trends within databases and leveraging this information to construct prognostic models [4]. Instead, it entails the systematic selection and exploration of data, and the construction of models utilizing expansive data reservoirs to unveil erstwhile undiscovered patterns [5].

Numerous individuals perceive data extraction as a facet of knowledge unearthing from databases, or KUIR. Data extraction can be deemed a pivotal stage in the knowledge unearthing process. The process of knowledge unearthing is depicted as an iterative progression of Data purification, Data integration, Data curation, Data transmutation, Data extraction, Pattern assessment, and Knowledge presentation.

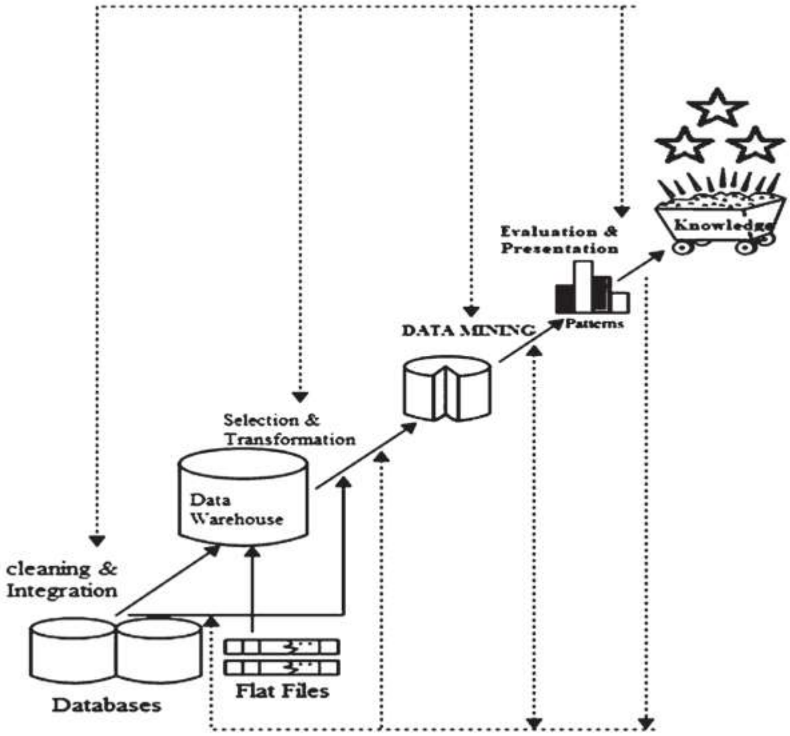


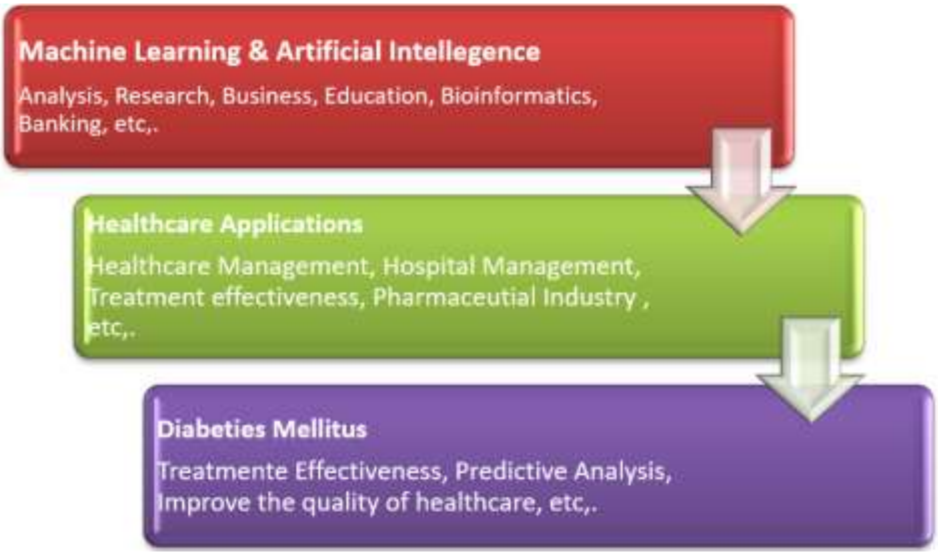
Figure 1. Data mining as a step in the process of knowledge discovery [1]

Data mining is gaining popularity in the field of medicine, as the sheer volume and complexity of healthcare data surpasses the capabilities of traditional analysis methods. By uncovering patterns and trends within vast and intricate datasets, data mining enhances decision-making processes. The healthcare industry stands to reap significant benefits from the applications of data mining.

In the realm of artificial intelligence, machine learning plays a crucial role. By employing statistical techniques, machine learning allows machines to learn and improve through experience. Rather than relying on explicit programming, these programs adapt their behavior based on the knowledge gained from data. This paradigm can be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves providing the machine learning program with both input data and corresponding labels. In this case, human intervention is necessary to label the data beforehand. On the other hand, unsupervised learning algorithms do not receive any labels. Instead, they autonomously uncover patterns and groupings within the inputted data. Lastly, reinforcement learning entails a computer program dynamically interacting with its environment, receiving feedback in the form of positive or negative reinforcement to enhance its performance.

Deep learning, on the other hand, revolves around the concept of acquiring multiple layers of representation and abstraction to make sense of various forms of data, such as images, sound, and text. This hierarchical learning approach enables a deeper understanding of complex information.


Machine Learning & Artificial Intellegence

Analysis, Research, Business, Education, Bioinformatics, Banking, etc,.

Healthcare Applications

Healthcare Management, Hospital Management, Treatment effectiveness, Pharmaceutial Industry , etc,.

Diabeties Mellitus

Treatmente Effectiveness, Predictive Analysis, Improve the quality of healthcare, etc,.

Figure 2. process diagram of proposed syste

Diabetes Mellitus: is a chronic ailment, arises when the pancreatic gland fails to produce an adequate amount of insulin or when the body's utilization of this vital hormone is ineffective. Insulin, a regulatory substance, is responsible for the management of blood sugar levels, also known as hyperglycemia, or elevated blood sugar. Uncontrolled diabetes often leads to severe harm to multiple bodily systems, particularly the nerves and blood vessels [WHO].

The classification of diabetes encompasses four distinct types: Type 1 diabetes, Type 2 diabetes, Pre diabetes, and Gestational diabetes. Type 1 diabetes mellitus occurs when the body is incapable of generating insulin, which is necessary for the conversion of food sugar into usable energy. This form predominantly affects children and young adults. On the other hand, Type 2 diabetes prevails as the most prevalent type, afflicting 90-95% of all individuals diagnosed with diabetes. Pre diabetes, sometimes referred to as impaired glucose tolerance, represents a milder manifestation of the disease. Gestational diabetes arises during pregnancy and poses a heightened risk for the mother's lifelong susceptibility to diabetes, as well as an increased likelihood of the infant developing obesity and diabetes. The International Diabetes Federation reports a current global estimate of 246 million individuals affected by diabetes, with projections indicating a surge to 380 million by 2025. India, in particular, witnesses the alarming escalation of diabetes mellitus, with a staggering 62 million diagnosed cases [9]

[10]. In 2000, India held the disheartening record for the highest number of individuals with diabetes mellitus, standing at 31.7 million, followed by China with 20.8 million and the United States with 17.7 million [11]. India's future remains uncertain, as it grapples with the potential burden that diabetes may impose on the nation.

Review of Machine Learning and Data Mining Methods in Diabetes Healthcare:

Ravi Sanakal et al embarked on a mission to devise a system that would aid physicians in medical diagnostics. Their study introduced a diagnostic FCM clustering technique and SVM using SMO to determine which method proves most effective in diagnosing diabetes mellitus. The FCM clustering approach exhibited an impressive accuracy of 94.3% and a positive predictive value of 88.57%. Conversely, the SVM method yielded a lower accuracy of 59.5%. These findings, though complex in nature, are relatively satisfactory given the intricate nature of diabetes detection [12].

In a similar vein, Nahla Barakat et al proposed the utilization of support vector machines (SVMs) for diabetes diagnosis. The authors also implemented an additional explanation module, known as the "black box model," which translates the SVM model into an intelligible representation of the diagnostic classification decision. SVMs emerged as a promising tool for diabetes prediction, boasting an accuracy of 94%, sensitivity of 93%, and specificity of 94%, all validated with real-life diabetes datasets [13].

G. Parthiban et al put forth a data mining approach aimed at extracting valuable correlations from attributes that may not directly indicate the class being predicted. Their study focused on predicting the likelihood of developing heart disease in diabetic patients based on diagnostic attributes. The authors demonstrated that it is indeed possible to diagnose the vulnerability of diabetic patients to heart disease with reasonable accuracy. Such classifiers could enable early detection of heart disease susceptibility in diabetic patients, prompting necessary lifestyle changes and preventing the onset of heart disease. SVM proved to be the optimal classification technique, exhibiting excellent predictive performance as evaluated through ROC curve analysis for both training and testing data. Consequently, the authors recommended the SVM model for the classification of diabetic datasets [14].

Berina Alic et al conducted a comparative analysis of machine learning techniques, specifically Artificial Neural Networks (ANN) and Bayesian Networks, for the classification of diabetes and cardiovascular diseases. ANN, particularly the multilayer feedforward neural network with Levenberg-Marquardt learning algorithm, emerged as the most widely utilized type. On the other hand, Na'ive Bayesian networks demonstrated the highest accuracy rates for the classification of diabetes and cardiovascular diseases, reaching 99.51% and 97.92% respectively. The observed mean accuracy of the network using ANN surpassed other models, indicating a higher likelihood of obtaining more precise results in diabetes classification [15].

The research conducted by Han Wu et al aimed to establish an effective prediction model for the high-risk type 2 diabetes mellitus (T2DM) group. This proposed model consisted of double-level algorithms, incorporating an improved K-means algorithm and logistic regression algorithms [16].

In their systematic efforts to design a prediction system for diabetes disease, Deepti Sisodia et al explored three different machine learning classification algorithms and evaluated them based on various metrics. The experiments were conducted using the Pima Indians Diabetes database [17].

Sajida Perveen et al developed a model with enhanced performance for classifying diabetic patients, employing bagging adaboost and J48 decision tree algorithms. The study utilized age groups from the Canadian population, extracting data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database. The evaluation of results indicated that the adaboost ensemble method outperformed both bagging and standalone J48 decision tree algorithms [18].

Saba Bashir et al proposed a framework for multilayer classification utilizing enhanced bagging and optimized weighting techniques. This framework was evaluated on five different disease datasets, including two diabetes datasets. The analysis of the output revealed that the HM-BagMoov ensemble framework achieved the highest accuracy, sensitivity, and F-Measure compared to individual classifiers. Furthermore, this framework surpassed state-of-the-art techniques in terms of accuracy, proving beneficial for disease classification and prediction based on disease symptoms [19].

Najmeh Hosseinpour et al applied several intelligent classifiers, such as Bayesian, Functional, Rule-based, Decision Trees, and Ensemble models, for diabetes diagnosis. The study utilized the PID dataset, with the results indicating that Functional and Ensemble classifiers exhibited the highest average F-measure and classification rate. Bagging with logistic core demonstrated top performance [20].

Dilip Kumar Choubey et al conducted a survey on various Soft Computing approaches, including Support Vector Machines, Ant Colony Optimization, Neural Networks, Fuzzy Logic, Genetic Algorithm, Rough Sets, and more, for the diagnosis of diabetes mellitus [21].

B.M. Patil et al proposed a hybrid prediction model (HPM) that utilized the simple K-means clustering algorithm to validate the chosen class label of given data and subsequently applied a classification algorithm to the result set. The final classifier model was built using the k-fold cross-validation method with the C4.5 algorithm. This hybrid model achieved a classification accuracy of 92.38% on the Pima Indians diabetes dataset [22].

K. Rajesh et al applied multiple classification algorithms to a diabetes dataset, analyzing the performance of each algorithm. The C4.5 algorithm exhibited a classification rate of 91% [23].

K. Rajesh and his colleagues conducted a comprehensive analysis on the Diabetes dataset, applying various classification algorithms. The performance of these algorithms was thoroughly evaluated. Notably, the C4.5 algorithm exhibited an impressive classification rate of 91% [23].

Moving on, Srideivanai Nagarajan and his team developed an expert system for diagnosing diabetes mellitus and assessing the risk levels among diabetic patients. They employed data mining techniques, specifically clustering and classification, to accomplish this task. The system successfully identified type-1, type-2, and gestational diabetes through the utilization of the Simple K-means algorithm [24].

In a similar vein, Ravi Sanakal and his associates designed a medical diagnosis system that employed a diagnostic FCM (Fuzzy Cognitive Map) as well as SVM (Support Vector Machine) with SMO (Sequential Minimal Optimization). Their objective was to determine which technique would be most effective in diagnosing Diabetes Mellitus. The SVM approach exhibited a rather disappointing accuracy of 59.5%, whereas the FCM model yielded outstanding results, boasting an accuracy of 94.3% and a positive predictive value of 88.57% [25].

Shifting gears, Manaswini Pradhan and his team developed an Artificial Neural Network (ANN) model for data classification. The ANN model was specifically evaluated for the task of pattern classification. The experimental studies conducted by the researchers confirmed the model's proficiency in performing pattern classification tasks. In fact, it outperformed other existing models and algorithms in terms of effectiveness [26].

In a different study, Abdullah A. Aljumah and his colleagues conducted predictive analysis on diabetic treatment, utilizing a regression-based technique. They employed the Oracle Data Miner as their data mining tool for predicting modes of treating diabetes. The SVM algorithm was specifically utilized for the experimental analysis, with the dataset being thoroughly examined to identify effective treatment methods for patients across all age groups [27].

V. Anuja Kumari and her team proposed an approach that utilized SVM with a Radial Basis Function (RBF) kernel for classification purposes. The data was trained using SVM, and the performance parameters, such as classification accuracy, sensitivity, and specificity, were found to be remarkably high for both SVM and RBF [28].

Krati Saxena and her associates employed the K-nearest neighbor algorithm for the diagnosis of diabetes. They presented accuracy and error rates for different mean values of K, specifically K=3 and K=5. The analysis revealed that many outputs from the test dataset matched the outputs from the training dataset, indicating a high level of accuracy. Notably, this accuracy was achieved by considering various features of the training dataset [29].

G. Parthiban and his team developed a model for predicting the likelihood of developing heart disease using attributes derived from diabetes diagnosis. The authors successfully demonstrated the feasibility of diagnosing heart disease vulnerability in diabetic patients with reasonable accuracy. The utilization of classifiers greatly facilitated the early detection of heart disease vulnerability in diabetic patients [30].

Lastly, J. Pradeep Kandhasamy and his colleagues compared the performance of various algorithms commonly used for predicting diabetes through data mining techniques. The authors employed machine learning classifiers, including J48 Decision Tree, K-Nearest Neighbors, Random Forest, and Support Vector Machines, to classify patients with diabetes mellitus. The performances of these algorithms were evaluated and compared based on accuracy, sensitivity, and specificity [31].

Conclusion:

In this comprehensive analysis, the authors delve into the intricate realm of Data Mining and Machine Learning, elucidating their profound implications in the realm of diabetes prediction and prognosis. The plethora of studies conducted in this domain predominantly revolves around the formulation of predictive models through the utilization of cutting-edge methodologies such as Support Vector Machines and K-Nearest Neighbors. These astute classification algorithms are meticulously designed to prognosticate the trajectory of the disease. Following a meticulous evaluation of the findings, it becomes evident that the support vector machine exhibits an exceptionally elevated level of accuracy, rendering it an optimal choice for the classification endeavor.

References:

[1]. Whiting, David & Guariguata, Leonor & Weil, Clara & Shaw, Jonathan. (2011). IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. Diabetes research and clinical practice. 94. 311-21. 10.1016/j.diabres.2011.10.029.

[2]. Sherwani, S. I., Khan, H. A., Ekhzaimy, A., Masood, A. & Sakharkar, M. K. Significance of hba1c test in diagnosis and prognosis of diabetic patients. Biomarker Insights11, BMI–S38440 (2016).

[3]. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4*4 million participants.

Lancet 2016; published online April 7. http://dx.doi.org/10.1016/S0140-6736(16)00618-8.

[4]. Seuring T, Archangelidi O, Suhrcke M. The economic costs of type 2 diabetes: A global systematic review. PharmacoEconomics. 2015; 33(8): 811–31.

[5]. IDF Diabetes Atlas, 6th ed. Brussels, International Diabetes Federation; 2013.

[6]. Nada Lavrac , “Selected techniques for data mining in medicine” , Artificial Intelligence in Medicine 16 (1999) 3–23

[7]. Frawley W, Piatetsky-Shapiro G, Matheus C. Knowledge discovery in databases: an overview. In Piatetsky-Shapiro G, Frawley W, editors. Knowledge discovery in databases. Menlo Park, CA: The AAAI Press, 1991.

[8]. HianChyeKoh and Gerald Tan,―Data Mining Applications in Healthcare, journal of Healthcare Information Management – Vol 19, No 2.

[9]. Mohammed Ali Shaik and Dhanraj Verma, (2020), Enhanced ANN training model to smooth and time series forecast, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022038, doi.org/10.1088/1757-899X/981/2/022038

[10]. Kincade, K. (1998). Data mining: digging for healthcare gold. Insurance & Technology, 23(2), IM2-IM7.

[11]. Milley, A. (2000). Healthcare and data mining. Health Management Technology, 21(8), 44-47

[12]. Mohammed Ali Shaik, Dhanraj Verma, P Praveen, K Ranganath and Bonthala Prabhanjan Yadav, (2020), RNN based prediction of spatiotemporal data mining, 2020 IOP Conf. Ser.: Mater. Sci. Eng. 981 022027, doi.org/10.1088/1757-899X/981/2/022027

[13]. Christy, T. (1997). Analytical tools help health firms fight fraud. Insurance & Technology, 22(3), 22-26

[14]. Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers is an imprint of Elsevier., 500 Sansome Street, Suite 400, San Francisco, CA 94111, ISBN 13: 978-1-55860-901-3

[15]. S.Yamini , Dr.V.Khanaa , Dr.Krishna Mohantha - A State of the Art Review on Various Data Mining Techniques, International Journal of Innovative Research in Science, Engineering and Technology, Vol. 5, Issue 3, March 2016

[16]. Mohammed Ali Shaik and Dhanraj Verma, (2020), Deep learning time series to forecast COVID-19 active cases in INDIA: A comparative study, 2020 IOP Conf. Ser.:Mater.Sci.Eng. 981 022041, doi.org/10.1088/1757-899X/981/2/022041

[17]. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., pp. 37–54, 1996.

[18]. J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, “Emerging information technologies for enhanced healthcare,” Comput. Ind., vol. 69, pp. 3–11, 2015.

[19]. N. Wickramasinghe, S. K. Sharma, and J. N. D. Gupta, “Knowledge Management in Healthcare,” vol. 63, pp. 5–18, 2005

[20]. . Mohammed Ali Shaik, “Time Series Forecasting using Vector quantization”, International Journal of Advanced Science and Technology (IJAST), ISSN:2005-4238,Volume-29,Issue-4 (2020), Pp.169-175.

[21]. Shortliffe, EH.,Perrault, LE., (Eds.). Medical informatics: Computer applications in health care and biomedicine (2nd Edition). New York: Springer, 2000.

[22]. Denis Rothman, “Artificial Intelligence by Example””, Ingram short title (2018),1788990544,50-250

[23]. Nick Bostrom,“Superintelligence: Paths, Dangers, Strategies”, Oxford University Press, 2014, ISBN 0199678111, 9780199678112

[24]. Shai Shalev-Shwartz, Shai Ben-David “Understanding Machine Learning”, Cambridge University Press,United States of America, ISBN 978-1-107-05713-5

[25]. Y. LeCun, Y. Bengio, G. Hinton, Deep learning Nature, 521 (7553) (2015), pp. 436-444

[26]. Joshi SR, Parikh RM. India - diabetes capital of the world: now heading towards hypertension. J Assoc Physicians India. 2007;55:323–4

[27]. Kumar A, Goel MK, Jain RB, Khanna P, Chaudhary V. India towards diabetes control: Key issues. Australas Med J. 2013;6(10):524–31.

[28]. Mohammed Ali Shaik, “A Survey on Text Classification methods through Machine Learning Methods”, International Journal of Control and Automation (IJCA), ISSN:2005-4297,Volume12,Issue-6

[29]. Kaveeshwar SA, Cornwall J. The current state of diabetes mellitus in India. Australas Med J. 2014;7:45–8”

[30]. Global report on diabetes. WHO Library Cataloguing-in-Publication Data, ISBN 978 92 4 156525 7.

[31]. Dilip Kumar Choubey, Sanchita Paul and Joy Bhattachrjee , “Soft Computing Approaches for Diabetes Disease Diagnosis: A Survey”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 9, Number 21 (2014) pp. 11715-11726