Detecting Diabetes Early and Classifying Risk with Inter0pretable Deep Learning Models

[1]Mohmad Ahmed Ali, [2]Dr. Hiren Dand

[1]Research scholar, Research scholar, Shri Jagdishprasad Jhabarmal Tibrewala University, Vidyanagri, Jhunjhunu, Rajasthan.

[2]Associate Professor, Dep of CSE, Shri Jagdishprasad Jhabarmal Tibrewala University, Vidyanagri, Jhunjhunu, Rajasthan

ABSTRACT

This study presents a novel approach to early detection and risk classification of diabetes using interpretable deep learning models, specifically the Deep Learning Multi-Layer Neural Network (DLMNN) framework. The DLMNN outperforms existing methodologies in prediction accuracy due to enhanced data preprocessing techniques and the MKMC algorithm for diabetes diagnosis. The Deep Learning-based Risk (DSR) classifier streamlines the risk analysis process, providing a detailed risk assessment of patients and quantifying potential risk levels associated with diabetes. The DSR classifier demonstrates the system's ability to classify individuals based on their likelihood of developing diabetes, supporting healthcare professionals in prioritizing high-risk patients and contributing to personalized healthcare strategies. The proposed interpretable deep learning framework for early diabetes detection shows significant advancements in accuracy and computational efficiency compared to traditional methods. The findings emphasize the importance of integrating interpretability into deep learning models, ensuring that predictions made can be understood and trusted by healthcare practitioners. This study lays the groundwork for further exploration into the application of deep learning technologies in diabetes management and emphasizes the potential for these models to improve patient outcomes through early diagnosis and tailored interventions.

KEYWORDS: DSR, DKMN, Diabetic and Deep Learning

I. INTRODCUTION

Our proposed approach to this study combines traditional machine learning with deep learning techniques. There are several separate phases to comprehensive method. Partitioning whole dataset into training and testing groups is first stage in Machine Learning. Synthetic Minority Oversampling Technique, or SMOTE, was used to improve training dataset such that two classes were evenly distributed. Because of this, more samples were gathered from under-represented instances, which improved their representation. We will next normalize attributes using min-max scaling algorithms. Perhaps a fairer distribution of traits will emerge from process. In this case, output is proportional to sum of all characteristics. Identifying most important properties of dataset was accomplished using a chi-square feature selection approach. Improving model performance while optimizing computation efficiency relies on feature selection.

Partitioning of dataset will follow chi-square test. You may expect to see 50% of dataset used for training and 50% put aside for validation attempts. Supervised machine learning techniques were then used to preprocessed training dataset. algorithm's capacity to shed light on interconnections between diabetes risk variables and its distinctive features served as guiding principles throughout algorithm selection process. Each method's efficacy was evaluated using a comprehensive set of evaluation criteria, including as F1-score, accuracy, precision, and recall. To improve model's performance, we tweaked its hyperparameters using validation dataset and used research methods. To begin, we utilized paper-specified hyperparameters for each model; thereafter, we used a manual tuning technique to gradually adjust values such that they conformed to recommendations [14]. We used area under receiver operating characteristic (ROC) curve, F1 score, accuracy, precision, and recall to evaluate our model's performance. In comparison to hyper-parameters derived from research, personal hyper-tuning minimizes probability of overfitting and significantly increases accuracy. best way to forecast diabetes in this specific case may be found by analyzing data. Following is a sequential description of method in figure.

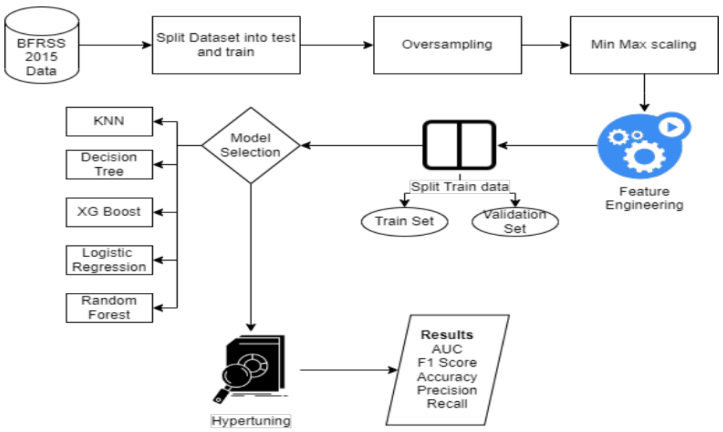

Figure 1: Flow Chart of Deep Learning Approach

In deep Learning approach, initial phase is dividing overall dataset into its relevant testing and training subsets. I have used stratify parameter during splitting of dataset to ensure that class distribution in original dataset is maintained in both training and testing sets. Then, oversampling was performed on training dataset to balance data for both classes using SMOTE (Synthetic Minority Oversampling Technique) analysis, which would allow generation of new samples based on minor instances to amplify their existence. Min-max scaling would be used next to standardize range of features. This would enable uniform distribution of attributes; these features would contribute equally to model. After completing min-max scaling, training data would be split into a train and validation set to perform deep learning algorithms. Then, deep learning algorithms were applied to preprocessed dataset. Each algorithm is chosen for its specific characteristics and potential to model relationships between diabetes risk factors. Through comprehensive experimentation, we evaluate performance of each algorithm on various evaluation metrics such as accuracy, precision, recall, F1-score, and area under receiver operating characteristic (ROC) curve. comparison of results allows us to identify most effective algorithm for diabetes prediction in our specific context. diagram below shows flow of methodology

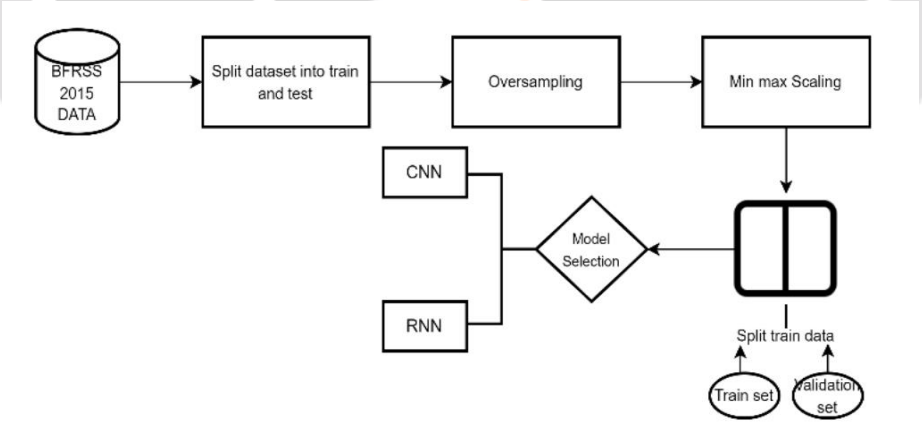

Figure 2: Flow Chart of Deep Learning Approach

II. Training and Validation Sets

To develop robust machine learning models, it is essential to utilize a synthetic dataset that aligns with specific problem statement. This will enhance model's ability to understand fundamental patterns and relationships. Upon training on 90% of dataset, which comprised both labelled and unlabeled samples, machine learning models exhibited a significant degree of accuracy in identifying relationships between traits and target variable of diabetes status. To minimize disparity between training and testing datasets, a sequence of modifications was applied to model's internal parameters, such as weights and thresholds, following each enhancement [4]. model's performance and its capacity to efficiently handle novel, unseen input can solely be assessed through validation dataset. We aim to evaluate model's performance by examining its learnt patterns across various scenarios. This is essential as it can substantially reduce likelihood of overfitting, a phenomenon where a model becomes overly specialized to training dataset, resulting in suboptimal performance on unseen data. A validation dataset could

potentially facilitate alignment of two types of datasets. dataset utilized for training was segmented, allocating 10% of data to establish validation set for this model.

Validation Dataset Hyper-tuning

Although effects may not be immediately seen in training datasets, validation dataset is crucial for studying parameter optimization and has a major influence on how well model's function. We maximized performance of all models to get optimal outcomes in novel, untested scenarios by modifying hyperparameters based on performance metrics from validation dataset. In order to boost model's overall performance, 10% of dataset tuples will be used as a validation set. Health prevention efforts are being greatly influenced by current emphasis on disease prediction using data mining. Presently, diabetes stands as foremost issue in world health. There are a number of methods that may be used to forecast likelihood of diabetes and its complications. At this time, there is a lack of accuracy in evaluation of relative dangers of diabetes and its impacts on various human organs. We propose a method to enhance diabetes prediction algorithms by reducing their current drawbacks.

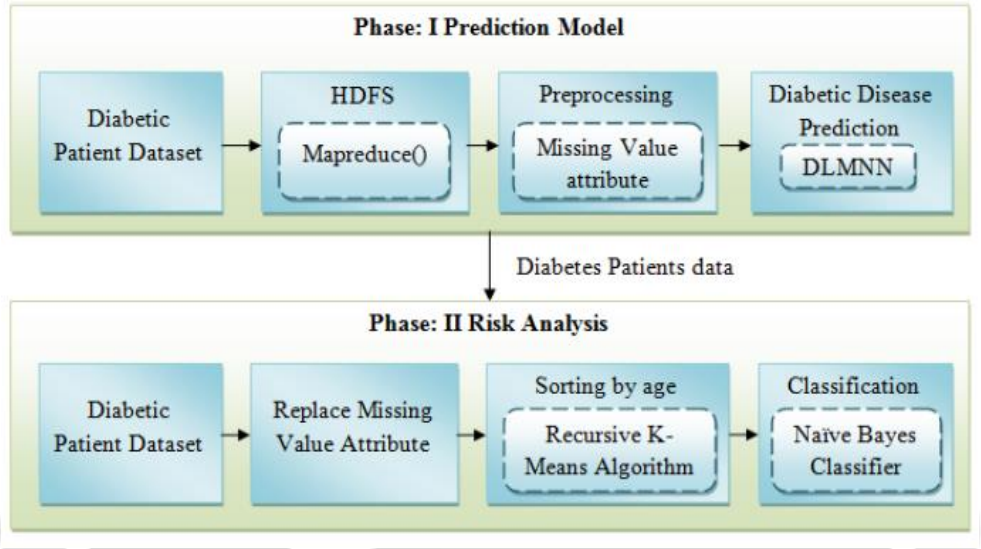

Figure 3 System Architecture

III. BDIABETIC PATIENT DATASET

The dataset consists of 768 patient records with '1' target class and 8 attributes of diabetes disease. Redundant data is eliminated using HDFS, which aids in creating replicated data due to increasing availability of data during node failure. Hadoop MapReduce underpins data deduplication and provides essential support for it. It removes unnecessary subfiles (sometimes called chunks, blocks, or extents) while reducing storage duplication to a minimum. Reduced storage requirements, decreased network traffic, and improved scalability are just a few of the many benefits of data deduplication.

With the help of MapReduce, entities in large datasets can be successfully resolved by partitioning the dataset into blocks that include records that are comparable to each other. Two MapReduce algorithms are used for entity identification: load balance and similarity computations, and occurrence counting. Dedoop can easily remove duplicate models by linking given workflows to appropriate execution units within the component library. Each JobCon contains fundamental MapReduce parameters, and the component library and newly created JobCon files are uploaded to the Hadoop cluster by Dedoop.

The study aims to help in fighting against diabetes by collecting samples from a large database using strict selection criteria. All women in the research are adults from the Pima Indian tribe. Medical predictors are a collection of independent factors that make up each dataset, with Outcome serving as the only dependent variable. The correlation between body mass index (kg/m²) and heritability is strong, and hereditary factors are responsible for almost half of all obesity cases.

The Hadoop Dispersed Record System (HDFS) is strengthened with stored data, enabling quick information exchange between hubs. MapReduce task is considered a framework that reads and writes functions to handle large volumes of data simultaneously, simultaneously, and large groups of commodity hardware clusters. It provides analytical capabilities for analyzing data and a program pattern for parallel computing that relies on programming language like Java.

MapReduce task reduces data handling complexity by allowing scaling several computing nodes data processing. Breaking down input data processing into subtasks like mappers and reducers is a complex problem, and MapReduce model offers simple scalability. Distributing tasks across nodes is made easy using MapReduce architecture, which implements Sort or Merge based on distributed computing. Input is fragmented into numerous chunks, and every single data chunk is implemented in different nodes.

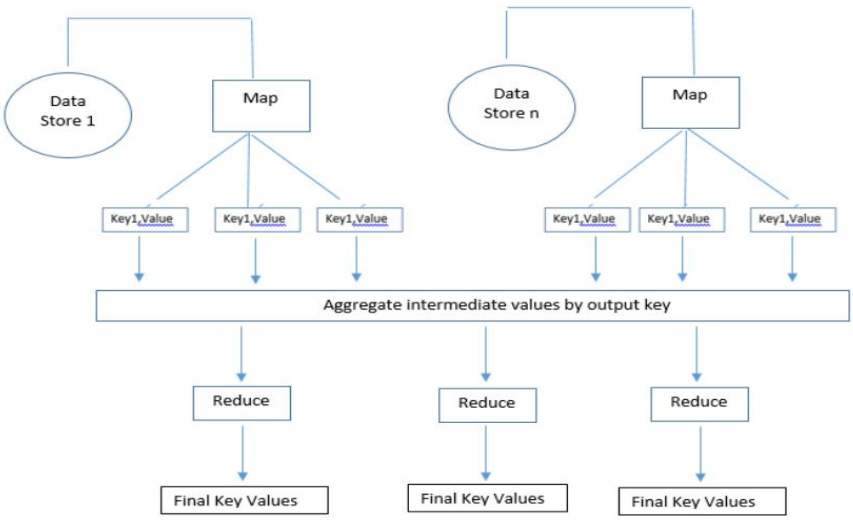

Figure 4 Map Reduce Architecture

IV. LOGISTIC REGRESSION

Logistic regression is a crucial tool in machine learning classification, as it evaluates independent factors that impact a single result variable. It assumes predictor independence and rely only on binary predicted variables, not taking missing data into consideration. Decision trees are robust data classification mechanisms that can process both numerical and categorical data types, offer a clear and interpretable structure, and require minimal data preprocessing efforts.

Fuzzy logic is a form of many-valued logic that deals with approximate reasoning rather than fixed and exact values, allowing for a more nuanced approach to problem-solving and decision-making. Random forest classifiers enhance prediction accuracy and mitigate overfitting by employing a meta-estimator to construct multiple decision trees from distinct sub-samples of datasets, subsequently averaging their outputs. They demonstrate superior performance compared to decision trees while effectively mitigating the risk of overfitting.

AdaBoost is a pioneering method established for enhancing performance in binary classification tasks and serves as an optimal foundation for acquiring knowledge in the field of advertising. AdaBoost is often used for shallow decision trees and is often used for shallow decision trees. The parameters LBXSOSSI (osmolality) (mmol/kg) and LBXSNASI (sodium) (mmol/L) play a critical role in enhanced AdaBoost classification model. Key vital signs in the AdaBoost classification model include duration of existence, mean cell hemoglobin (pg), chloride concentration (mmol/L), waist circumference (cm), and LBXSCLSI. These acronyms represent key vital signs such as systolic blood pressure (mm Hg), direct HDL cholesterol (mg/dL), gamma glutamyl transferase (U/L), PHDSESN_1.0, indicating that examination session occurs in morning, and LBDHDD. In summary, logistic regression, decision trees, fuzzy logic, and AdaBoost are essential tools for machine learning classification and prediction accuracy. These methods help address the challenges of real-time forecasting and provide a more accurate and reliable classification of data.

Gradient boosting is an algorithmic approach used in machine learning for classification and regression. It involves sequential and additive training of multiple models, with AdaBoost focusing on high-weight data points. The absence of a diabetes diagnosis from a physician is a critical aspect of the enhanced Gradient Boosting classification model.

XGBoost, an ensemble method in machine learning, utilizes decision trees trained through a gradient boosting framework. Artificial neural networks often demonstrate superior performance compared to alternative methods.

The XGBoost classification model identifies DIQ010_2.0 as a significant feature: there has been no prior mention of diabetes by a physician.

The proposed architecture includes data collection, preprocessing, model training, performance evaluation, and ultimately facilitating predictions regarding categorization of diabetic diet quality. The NHANES dataset is structured in a concise file format that focuses on a specific subject area for each survey cycle. Data is collected from HDFS and undergoes preprocessing according to industry standards before being exported in.csv format.

Diabetic individuals are classified into three categories using laboratory results, demographic data, clinical examination outcomes, nutritional practices, and responses to questionnaires. The pre-processed NHANES dataset serves as the foundation for constructing a rule-based machine learning method integrated within PCUDD. Accuracy is a key metric for evaluating accuracy in classification tasks. To achieve optimal performance, accuracy should ideally approach a value of 1, with a precision score of 1 achieved when numerator and denominator are equivalent. Recall, also known as sensitivity and true positive rate, signifies the presence of an effective classifier. A recall value of 1 or greater signifies the presence of an effective classifier. The F1-score is a robust metric for assessing a model's efficacy in classification tasks, combining both recall and precision into a single evaluative statistic. The two-score is computed as two times the discriminant of precision and recall, divided by their sum.

The F1 Score achieves a value of 1 when both precision and recall are maximized at 1. Optimal F1 score is achievable solely under conditions of exceptionally high recall and accuracy. A more intricate metric than accuracy, F1 score signifies harmonic mean of recall and precision.

V. PERFORMANCE ANALYSIS OF MKMC ALGORITHM

Here, proposed MKMC's performance in DM prediction is contrasted with already available techniques like Recursive K-means Clustering (RKMC) and Fuzzy C-means clustering algorithms.

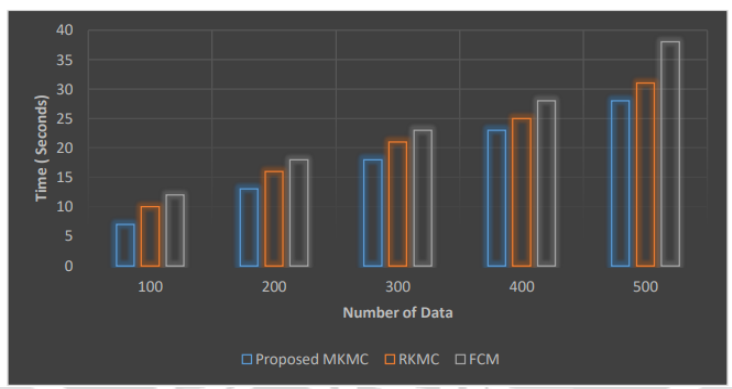

Figure 5 Clustering time comparison

Figure 5 delineates performance of proposed MKMC and existing RKMC and FCM. Computational time varies based on different numbers of data. For 100 data, proposed MKMC has consumed 9.011s for completing process but prevailing RKMC and FCM have taken 11.34s and 14.01s, which is high compared to proposed system. Correspondingly, for all remaining data values, existing techniques takes high time compared to proposed MKMC.

PERFORMANCE ANALYSIS OF DSR CLASSIFIER

The proposed DSR classifier used in risk analysis is contrasted with proposed DLMNN and already available approaches, such as Improved KMeans algorithm (IKMC), Deep Convolutional Neural Network (CNN), and Artificial Neural Network (ANN) approach in respect of precision, FMeasure, accuracy, recall

Fig 6 Performance Comparison of proposed DSR classifier with Proposed DLMNN and existing classifiers

Metrics	Proposed DLMNN	Proposed DSR	IKMC	CNN	ANN
Precision	97.82	97.02	93.54	92.47	91.37
Recall	97.82	93.82	90.78	90.14	88.04
F-Measure	97.82	94.01	92.11	91.37	87.82
Accuracy	96	94.9	93.24	92.13	89.98

The research work contributes to overcome issues addressed by DLCNN where slow convergence is considered as an issue. performance analysis of various existing model revealed no improvements when applied algorithm on Hungarian data set. Hence, DSR algorithm is used with a to identify whether it outperforms other existing algorithms in terms of accuracy.

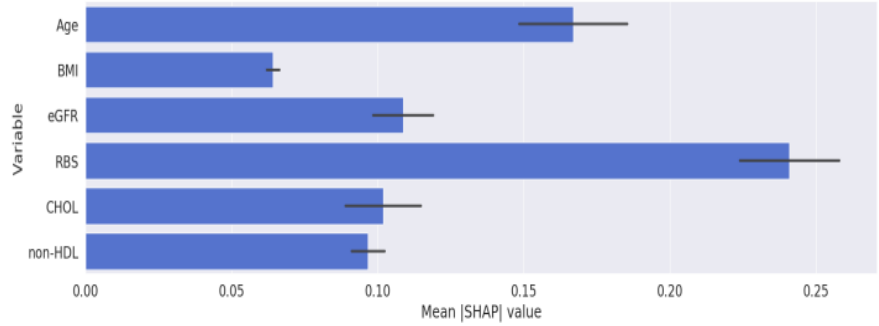

Figure 7: Relative order of importance of predictors for MLP model trained with longitudinal data using SHAP.

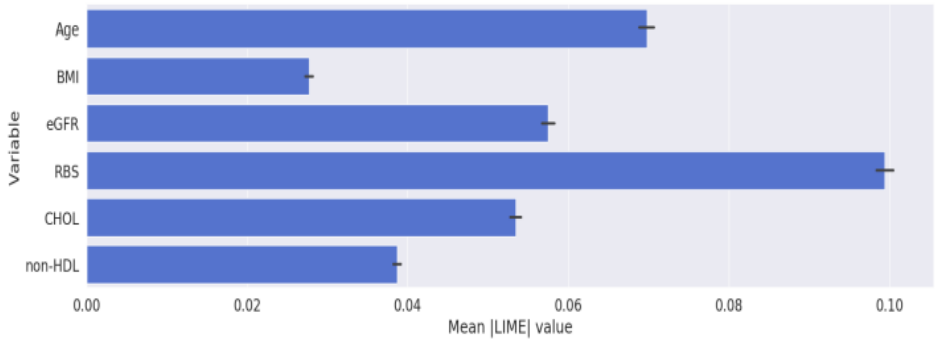

Figure 8: Predictor significance for LIME-trained MLP model with longitudinal data

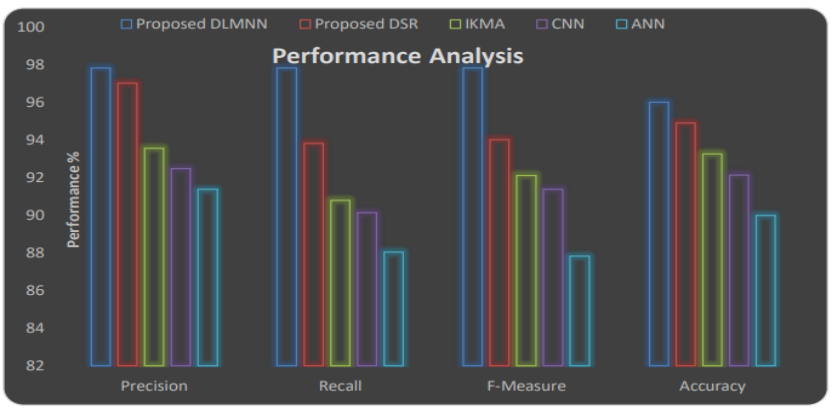

Figure 9 Performance Analysis of Proposed DLMNN, DSR with Existing Techniques

CONCLUSION

The Deep Learning Multi-Layer Neural Network (DLMNN) is a novel approach to diabetes prediction, offering superior accuracy compared to existing methods like IKMC, CNN, and ANN. The system uses the MKMC algorithm for diabetes diagnosis and the innovative DSR classifier for risk analysis, minimizing complexity while providing precise risk level assessments. The DSR classifier not only improves predictive outcomes but also offers a more comprehensive understanding of diabetes risk levels. The proposed approach shows great promise in enhancing diabetes prediction accuracy and efficiency, paving the way for more effective patient management and intervention strategies. The system helps physicians in early prediction of diabetes and risk levels, providing high performance in recall, accuracy, precision, and computational time. Future enhancements include implementing techniques on large amounts of data, reducing redundant data storage, and testing the effectiveness of data shrink techniques. The accuracy estimation can be analyzed using techniques with large data sets with increased attribute values, providing a more accurate early diabetic prediction model for large datasets.

REFERENCES

1. Andrew S Greenberg, Martin S Obin, ―Obesity and role of adipose tissue in inflammation and metabolism‖, Am J Clin Nutr; 2006; 83: 4615- 4655.
2. Banu Turgut Ozturk, Banu Bozkurt, Effect of serum cytokines and VEGF levels on diabetic retinopathy and macular thickness, Mol Vis; 2009; 15:1906–1914.
3. Chantelau E, Kimmerle R, Meyer-Schwickerath R, Insulin, insulin analogues and diabetic retinopathy, Arch Physiol Biochem; 2008; 114: (1): 54-62.
4. Ebru Nevin Cetin, Yunus Bulgu, Seyfullah Ozdemir, Senay Topsakal, Fulya Akin, Hulya Aybek, Cem Yildirim, Association of serum lipid levels with diabetic retinopathy‖, Int J Ophthalmol; 2013; 6 (3): 346-349.
5. Ramchandran, India Diabetes Research Foundation, Socio-Economic Burden of Diabetes in India‖, SUPPLEMENT OF JAPI JULY 200755.
6. Gaede P, Vedel P, Parving HH, Pedersen O, Intensified multifactorial intervention in patients with type 2 diabetes mellitus and microalbuminuria: steno Type 2 randomised study, Lancet; 1999; 353(9153):617-22.
7. Prasadu Peddi, & Dr. Akash Saxena. (2016). STUDYING DATA MINING TOOLS AND TECHNIQUES FOR PREDICTING STUDENT PERFORMANCE. International Journal Of Advance Research And Innovative Ideas In Education, 2(2), 1959-1967.
8. J T Gillow, J M Gibson, P M Dodson, Hypertension and diabetic retinopathy- what's story‖, British Journal of Opthalmology; 1999; 83, 1083-1087.
9. Petersen KF, Shulman GI, Etiology of insulin resistance‖, Am J Med; 2006; 119: S 10-6.
10. Zhang, Q. L., & Yang, Y. B. SA-Net: Shuffle attention for deep convolutional neural networks. Computer Vision and Pattern Recognition (2021).
11. Prasadu Peddi, & Dr. Akash Saxena. (2015). The Adoption of a Big Data and Extensive Multi-Labled Gradient Boosting System for Student Activity Analysis. International Journal of All Research Education and Scientific Methods (IJARESM), 3(7), 68-73.