*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 26, Issue 4, Ser. 2 (Jul. – Aug. 2024), PP 30-37*
*www.iosrjournals.org*

# Efficient Feature Selection Technique for Classification of Diabetic

[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***
*An innovative feature selection method for improving diabetic retinopathy (DR) model classification accuracy is introduced in this research. A high-dimensional dataset is created while evaluating a wide variety of patient health characteristics for a diabetes diagnosis, which poses a challenge to computing efficiency and model performance. The goal of the study is to minimise the dimensionality of the dataset while keeping the most useful features for classification tasks by using an optimised feature selection technique. In order to choose the smallest possible collection of characteristics that maximises classification accuracy, this method combines cutting-edge machine learning algorithms with statistical analysis. Results from experiments conducted using diabetes datasets that are accessible to the public indicate that the suggested method works, with significant gains in classification accuracy and computing performance. As a more efficient and precise diagnostic tool for diabetes categorisation, this approach shows promise for use in clinical settings. In terms of sensitivity and accuracy measures, the ensemble model performs better than individual machine learning methods when it comes to diabetic retinopathy categorisation. This study presents Enhanced Lasso, a new technique for feature selection that minimises feature redundancy and achieves better performance than state-of-the-art approaches when used during the ensemble learning phase.*

***KEYWORDS:*** *Diabetes mellitus (DM), diabetic retinopathy, RNN and CNN.*

## I. INTRODUCTION

Diabetes diagnosis can significantly impact one's life, affecting relationships and overall well-being. Maintaining emotional well-being is crucial for maintaining effective diabetes treatment, as it affects both physical and mental health. Mental disorders like depression and anxiety are undertreated among diabetics due to their reduced rates of diagnosis. At least 45% of cases of psychiatric illness and emotional impairment are never known because a patient has diabetes. Diabetes recovery teams must be mindful of the relationship between mind and body and take into account the body's overall well-being. Diabetes is a frequent source of mental health problems for those who don't pursue medical assistance. Individuals with too-high or too-low blood sugar levels are often asymptomatic, and they may experience discrimination and reluctance to express distress in public environments. Diabetes is caused by excessive sugar in the blood, which affects all brain processes, including thought, judgment, feeling, and behavioral processes. People with diabetes have a higher chance of experiencing depression, which is linked to the potential connection between stress and brain activity fluctuations. Improperly regulated diabetes may result in signs similar to depression. The likelihood of diabetes triggering dementia is higher and well-established, with chronic tension built up by elevated glucose on the liver potentially having detrimental effects on brain health. Diabetes also constricts blood arteries, contributing to hypertension, which induces decreased blood pressure and eventually causes brain injury over time. Delirium, often confused with diabetes, is often not properly regulated and can lead to long-term neurological decline and death. It is a matter of concern because delirium sometimes goes undetected and there is a short window to cure it within a person's lifetime. Families have been around for hundreds of thousands of years, but definitions have evolved over space and time. Marriage refers to a man and a woman, while spouse refers to an individual of opposite sex who is someone's husband or wife. An upset marital relationship can negatively impact physical and mental health, and life quality of life. A lot of time is spent in one day, and couples come into close contact, showing everything they have.

Diabetic retinopathy (DR) is a prevalent condition in industrialized nations, affecting approximately 10% of individuals diagnosed with diabetes. This condition adversely affects visual acuity and can lead to complete vision loss after 15 years. The global diabetes population is projected to rise significantly from 171 million in 2000 to 366 million by 2030, with the prevalence of diabetes across all age categories increasing from 2.8% in 2000 to 4.4%.
Patients must prioritize examination of DR to preserve their vision effectively and promptly. Automated diagnostic technologies significantly decrease labor intensity involved in screening processes, making it feasible to transition

from traditional manual methods for diagnosing diabetic retinopathy. Christodoulides et al. (2017) emphasized the necessity of creating an autonomous system capable of distinguishing between normal and abnormal cases to improve screening programs for the entire community.

Automated methodologies are available to evaluate morphological changes in retinal vasculature, which could serve as potential indicators of underlying retinal or cardiovascular pathology. Diabetic retinopathy is a prominent contributor to visual impairment and blindness. Artificial intelligence has gained significant attention due to its interdisciplinary, multidisciplinary, and transdisciplinary applications within Computer Science. The field of healthcare has undergone significant transformation due to innovations in digital image processing and extensive capabilities offered by AI.

Medical imaging technology has led to a significant rise in volume of medical images captured, facilitating the creation of an extensive database for analytical purposes and disease forecasting. The application of artificial intelligence holds significant promise for enhancing healthcare research methodologies. Recent advancements in medical technology include prediction of protein three-dimensional structures for drug development, identification of previously unknown relationships between viruses and stem cell categorization, and ability to anticipate the onset of chronic diseases.

Diabetic retinopathy is a critical pathological condition associated with diabetes mellitus, characterized by progressive degeneration of retinal structures, substantial or complete advanced stages. It affects 29 million individuals in developed countries, particularly within the age range of 25 to 74 years, with DR affecting 33% of this demographic.

## II. METHODOLOGY AND DATASETS

The datasets MESSIDOR, IDRID and APTOS were used for experimental analysis of this proposed model
The proposed model is implemented in three phases: Pre-processing, Feature selection and Ensemble based classification as shown in figure 4.1.

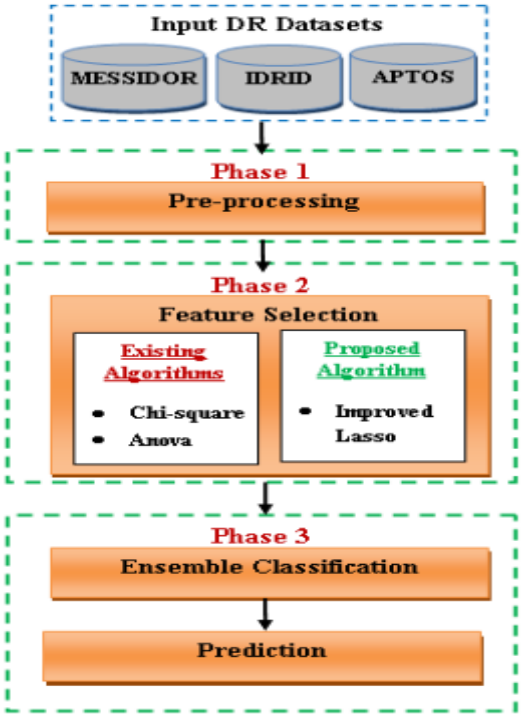


**Figure 1 System Architecture – Feature Selection**

**Pre-processing**

Machine learning algorithms play a crucial role in the success or failure of their pre-processing step. The DIDRO hybrid approach for data preparation consists of three basic steps: data imputation, data reduction, and one-hot encoding. Datasets for Diabetic Retinopathy (MESSIDOR, IDRID, and APTOS) were preprocessed using this technique to extract relevant features. Feature selection is the main goal of machine learning algorithms, with conventional feature selection approaches based on relative relevance of various qualities. This study proposes an Improved Lasso (IL) algorithm for selecting significant features from the DR dataset, which works by combining weights to zero and shrinking survivors toward equality. Regularization procedures can lead to enhanced prediction accuracy, reduced variability, increased model interpretability, and a diminished risk of overfitting. The Lasso (L1) penalty effectively identifies features with non-zero coefficients, and when aggregating absolute values of β, it is essential to comply with the upper limit, which is a predefined threshold. The penalty function in LASSO

causes coefficients to constantly shrink in the direction of zero, resulting in a sparse collection of variables with non-zero regression coefficients. When the Lambda value increases by i+1, it increases forecasting performance and interpretability of the regression model through sparsity. Ensemble learning creates a collection of independent learners who are combined with tactics such as bagging, boosting, and stacking. Ensemble-based models have recently gained popularity in usage for decision-making systems due to enhanced performance and better classification models. Two benefits of employing an ensemble-based model are performance and robustness. This suggested model would deliver the best answer, outperforming any single contributing model, and minimizes the spread or dispersion of model performance and forecast accuracy. Trees have a significant impact on machine learning settings, including classification and regression, in addition to many practical applications they offer. Decision trees are a helpful visual tool used in decision analysis and data mining and machine learning. The k-nearest neighbors (kNN) classifier is a non-parametric supervised machine learning technique that classifies objects based on their distance from one another and classifications of their immediate neighbors. In addressing classification and regression problems, a common methodology employed is the utilization of approach entails collecting multiple samples, organizing them according to their mean values, and applying regression algorithms to construct decision trees through a majority voting framework. Random Forest Algorithm exhibits notable capability to handle data sets comprising both continuous variables, often encountered in regression tasks, and categorical variables, typically found in classification scenarios. In the realm of categorization tasks, it demonstrates exceptional performance.

**Algorithm Improved Lasso**
Input: DR Datasets (MESSIDOR and IDRID)
Output: Selected Features
begin
Phase-I: Preprocessing
Step 1. Import Scikit learn library
Step 2. Impute missing values and replace with NAN values
Step 3. Apply one-hot encoding to convert categorical into binary vectors
Step 4. Reduce high dimensional to low dimensional DR datasets.
Phase-II: Feature Selection
Step 5. X and Y are training set that contains number of samples
Step 6. Determine each feature's mutual information value
Step 7. mutual information values of features are sorted in ascending order
Step 8. Use improved lasso method to update information
Phase- III: Classification using Ensemble Classification
Step 9. Perform classification of DR based on selected features from phase II
End

Utilizing recursive K-Means clustering approaches, dataset is organized into clusters based on common properties, such age. data that has been clustered is then classified into different categories using Naive Bayes approach. Amputation risk, renal impairment, color vision impairment, and cardiovascular disease may all be better evaluated using new approach. likelihood of developing cardiovascular disease, color vision deficit, renal failure, or amputation during a three-year period must be quantified. In analytical phase, anticipated results of preliminary stage are tested. This document covers every single job that has to be done for this phase.

Amputation, color vision impairment, Hungarian demographics, and renal impairment datasets are all a part of proposed system. PIMA dataset, which pertains to female participants who have been diagnosed with diabetes, is merged with integrated dataset. Common characteristics, including gender and age, allow for smooth integration of datasets. Entity resolution in massive datasets was accomplished with help of MapReduce techniques. To limit analysis to just entities inside each cluster, blocking-based entity matching uses a semantic partitioning approach to group related data into clusters. Analysis counts instances in entity resolution jobs, whereas matching helps with load balancing and similarity metric calculations. It also uses a greedy load balancing technique, which is sorting matched workloads by size and then allocating them to least load-intensive tasks. Table 4.1 shows relative times of calculations, showing how processing of data sets may be optimized using deduplication techniques to save time.

**Fig 2 Computation time comparison before and after DE duplication**

| Method | Time | Prediction Accuracy |
|---|---|---|
| Before DE-duplication | O(n2) – 9.7 µs per loop | 94.32% |
| After DE-duplication | O(n) – 4.7 µs per loop | 96% |

Standard mean equation as elucidated in phase -1 is utilized to handle missing attributes in dataset. The attribute that is common among different datasets such as age is used to sort dataset. RKMC algorithm is applied on dataset to group. In RKMC algorithm, fixed partition level is achieved. K-Means calculation is utilized recursively on each segment in order to attain fixed level of partition. Then, each segment has labelled archives of a single class. Both unlabeled and labelled information are segregated using RKMC calculation. steps involved in attaining fixed partition level is depicted below,

Table 1 Cluster Analysis

| Cluster | Age | Diabetes features | Kidney failure features | Heart disease features | Diabetic retinopathy features | Amputation features |
|---|---|---|---|---|---|---|
| Cluster 1 | 1 – 20 | No. of times pregnant | Anemia | Chest pain type | Diameter of the optic disc | Aniseptic |
| | | Plasma glucose | Blood glucose random | Fasting blood sugar | Quality assessment | Limb |
| | | Blood pressure (BP) | BP | BP | MA detection | |
| Cluster 2 | 20 – 40 | Triceps skinfold thickness | Specific gravity | Cholesterol | | |
| | | Serum insulin | Albumin | Resting electrocardiographic results | | |
| | | BMI | Sugar | Maximum heart rate attained | | |
| Cluster 3 | 40 – 60 | Pedigree function | Red blood cells (RBC) | Exercise-induced angina | | |
| | | | Pus cell | ST depression | | |
| | | | Pus cell clumps | Slope of a peak exercise ST segment | | |
| | | | Bacteria | Numbers of major vessels | | |
| | | | Blood urea | Thal | | |

Table 1 reveals cluster performance. There are five cluster groups. In cluster-1 consists of people's records with age 1 - 20 (age), Cluster-2 is a collection of people's record with age 20-40, cluster-3 is a collection of people's record with 40-80, cluster-4 is a collection of people's record with 60-80 and cluster-5 is a collection of people's record with age 80-100. In this table, features of diabetes disease are depicted. Symptoms of diabetes mellitus include hypertension, low blood glucose, impaired gene expression, and low insulin levels in blood. Levels of albumin, fructose, specific gravity, and iron deficiency are indicators of renal function. Heart disease symptoms, including chest discomfort, are among most important warning signs. We also cover characteristics of diabetic retinopathy and amputation. properties that are supplied are used by dataset for processing activities that follow.

NB Classifier algorithm is utilized for classifying cluster values. Naive bayes classifiers assume that each feature is only condition on class Diabetes leads to an enormous figure of death and higher population affected by disease are failed to recognize their health condition at early stage. This chapter proposes a model that aids in predicting Diabetes disease and its risk level analysis at early stage. proposed system contains '3' phases, namely, DM prediction model, Risk analysis, and Early prediction model. In first phase, initially, dataset is preprocessed, and then required features are extricated from dataset. Shrinking size of dataset results in reduced computational time. Finally, disease is predicted using MKMC algorithm. In phase 2, proposed system minimizes complexity by using a novel DSR classifier for risk analysis. In phase 3, non-diabetic patient data is undergoing preprocessing, and then important features are extorted from dataset. Finally, potential level of patient against DM is predicted by EDR method. outcomes are examined and contrasted with existing techniques to confirm that proposed model provides better results.

The proposed method for diabetes mellitus prediction makes use of Pima Indian Diabetes Dataset. It was internet repository at UCI that provided PIDD dataset. Eight distinct data factors characterize 768 female medical records that make up PIDD dataset; a class of '0' indicates a poor diabetes result and a class of '1' indicates a good

one. In order to better understand patient's health state, proposed method integrates laboratory data from previous year with blood glucose test results.

Due to fact that data quality substantially impacts outcomes of predictive models, data pre-processing is crucial in disease forecasting. In order to make most of dataset, pre-processing step involves two parts. An integral part of procedure is dealing with data duplication and absence of suitable alternatives.

The recommended approach to predicting diabetes mellitus takes into account important factors such as number of pregnancies, blood sugar levels, diastolic blood pressure, triceps skinfold thickness, two-hour serum insulin, BMI, family history of diabetes, and chronological age.

Fig 3 Extracted Feature set

| Features | Extracted Feature Set |
|---|---|
| Patient's age | Patient's age (F1) |
| Body Mass Index (BMI) | |
| Family History of Diabetes | Body Mass Index (BMI) (F2) |
| Physical Stress | |
| Blood pressure | Fasting blood glucose test results (mg/dL) (F3) |
| Smoking | |
| polycystic ovary syndrome (PCOS) history | Oral glucose Tolerance test results (mg/dL) (F4) |
| Being of African-American, Native American, Latin American, or Asian-Pacific Islander descent | |
| Fasting blood glucose test results (mg/dL) | HbA1c examination (percent) (F5) |
| Oral glucose Tolerance test (mg/dL) | |
| HbA1c examination (Percent) | |

The MKMC algorithm is utilized for patient outcomes categorization as positive and negative class. Here, positive class represents data is categorized as patient affected by diabetes. Here, value is 1 represents positive outcome and 0 represents negative outcome. negative outcome represents that specific data is not affected by diabetes. proposed K-means clustering algorithm uses Mahanttan distance metric for distance calculation. So, this clustering algorithm is termed as Mahanttan based K-Means Clustering (MKMC) algorithm.

## III. EXPERIMENTAL RESULTS

The 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. 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.

Fig 4 Comparative Analysis – IDRID Dataset

| Feature Selection Algorithms | ML Classification Algorithms | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| Chi-square | Decision Tree (DT) | 89 | 90 | 88 |
| | k-NN | 90 | 91 | 89 |
| | **Random Forest (RF)** | **92** | **93** | **92** |
| Anova | Decision Tree (DT) | 86 | 86 | 85 |
| | k-NN | 88 | 87 | 87 |
| | **Random Forest (RF)** | **93** | **94** | **93** |
| Improved Lasso | Decision Tree (DT) | 88 | 89 | 86 |
| | k-NN | 94 | 93 | 84 |
| | **Random Forest (RF)** | **96** | **97** | **92** |

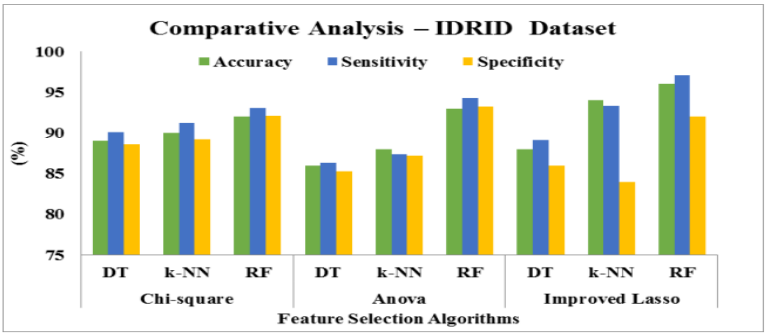


Figure 5 Comparative Analysis – IDRID Dataset

Fig 6 Performance Analysis – APTOS Dataset

| Feature Selection Algorithms | ML Classification Algorithms | Precision | Recall | F-Score |
|---|---|---|---|---|
| Chi-square | Decision Tree (DT) | 81 | 88 | 84 |
| | k-NN | 82 | 79 | 80 |
| | **Random Forest (RF)** | **89** | **93** | **90** |
| Anova | Decision Tree (DT) | 84 | 87 | 83 |
| | k-NN | 84 | 87 | 82 |
| | **Random Forest (RF)** | **90** | **94** | **91** |
| Improved Lasso | Decision Tree (DT) | 86 | 82 | 81 |
| | k-NN | 91 | 95 | 92 |
| | **Random Forest (RF)** | **94** | **91** | **90** |

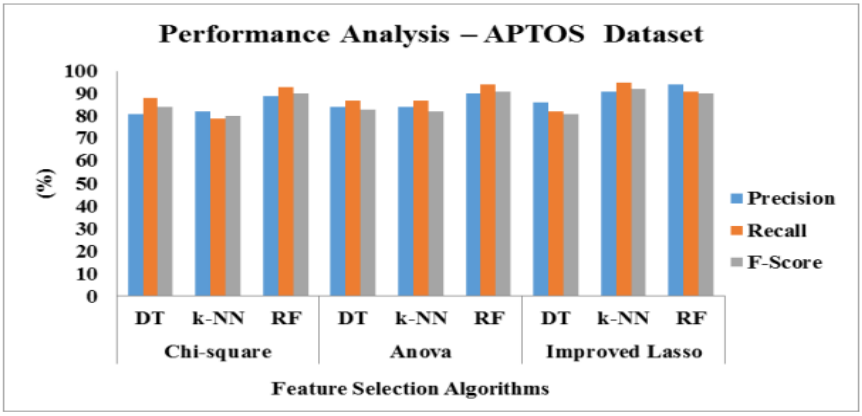


Figure 7 Performance Analysis – APTOS Dataset

From experimental results, it is observed that proposed Improved Lasso and Random Forest classifier produces higher Accuracy, Sensitivity, and Specificity than other feature selection classification algorithms with respect to APTOS datasets.

Fig 8 Comparative Analysis – APTOS Dataset

| Feature Selection Algorithms | ML Classification Algorithms | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| Chi-square | Decision Tree (DT) | 83 | 84 | 81 |
| | k-NN | 86 | 87 | 85 |
| | **Random Forest (RF)** | **90** | **93** | **90** |
| Anova | Decision Tree (DT) | 85 | 82 | 82 |
| | k-NN | 87 | 89 | 84 |
| | **Random Forest (RF)** | **92** | **90** | **85** |
| Improved Lasso | Decision Tree (DT) | 88 | 86 | 84 |
| | k-NN | 92 | 91 | 88 |
| | **Random Forest (RF)** | **94** | **95** | **92** |

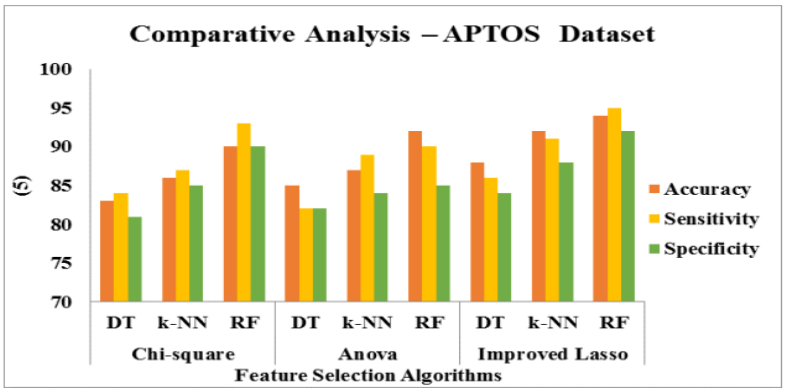


Figure 9 Comparative Analysis – APTOS Dataset

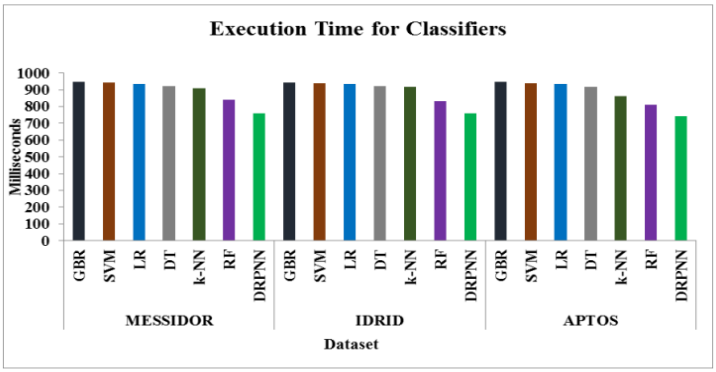


Figure 10 Execution Time for Classification Algorithms

Support Vector Machines (SVMs) have superior generalization capabilities compared to traditional multi-layer perceptron models. They choose a hyperplane that maximizes margin, allowing for better classification of items into well-defined categories. Machine learning techniques like one-against-one SVM, tree-structured SVM, and one-against-all SVM have been developed to identify clean or hard-labeled data. Upgraded architectures offer various training and decision-making approaches based on algorithms. Fuzzy SVM (FSVM) distributes membership over multiple classes for every observation.

A study using DRIVE and MESSIDOR databases was conducted to gather information on retina's non-dilated fundus. Out of 220 retinal input photos, 70 were considered normal, while 150 were classified as deviant. Thirty of the standard images were used during training, while the other forty were used for evaluating classifier techniques. Sixty of the erroneous photos were used in the training phase, while 90 were saved for later testing. A total of 145 photos, including 100 out-of-the-ordinary cases and 45 typical ones, were used to evaluate classifier methods.

## IV. CONCLUSION

In this paper, we have proposed Improved Lasso algorithm for feature selection using three DR datasets. These selected features are fed as input to classification stage where an ensemble learning-based classification method is implemented for classification of DR. experiments clearly show proposed Improve Lasso algorithm outperforms than existing feature selection algorithms. There is a new approach to diabetic retinopathy categorization called Diabetic Retinopathy Prediction using a Neural Network (DRPNN). Classification based on ensembles and regression were two of most well-known machine learning approaches used to assess suggested methodology. results were evaluated on several benchmark datasets, including MESSIDOR, IDRID, and APTOS. suggested DRPNN algorithm outperforms its rivals, according on experimental findings.

## REFERENCES

[1]. Alex Wright, A.C.Felix Burden, Richard B. Paisey, Carole A. Cull, Rury R.Holman, Sulfonylurea Inadequancyl, Daibetes Care; 2002; 25(2): 330- 336.

[2]. Chun Zhao, Weifang Wang, Ding Xu, Hui Li, Min Li and Fang Wang, ―Insulin and risk of Daibetic retinopathy in Patients from a meta-analysis of seven cohort studiesl, Diagn Pathol; 2014; 9: 130.

[3]. Gregory A N.ichols, Teresa A. Hillier, Jonathan B. Brown, ―Normal fasting plasma glucose and risk of Type 2 Daibetes diagnosisl, American Journal of Medicine; 2008; 121 (6): 519-524.

[4]. Donald S. Fong, Lloyd Aiello,Thomas W. Gardner, George L. King, George Blankenship, Jerry D. Cavallerano, et al., lRetinopathy in Diabetesl, Diabetes care; 2004; 27: s84-s87.

[5]. Wetterslev, J.; Jakobsen, J.C.; Gluud, C. Trial sequential analysis in systematic reviews with meta-analysis. BMC Med. Res. Methodol. **2017**, 17, 39.
[6]. Harol D A. Kahn and Robert F. Bradle, ―Prevalence of diabetic retinopathy-Age, sex, and duration of diabetes‖, Brit J Ophthal; 1975; 59: 345.
[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]. Y. Cho & L. K. Saul, Kernel Methods for Deep Learning, Proceeding of Neural Information Processing Systems (NIPS), pp. 342-350, 2009.
[9]. Gregory A N.ichols, Teresa A. Hillier, Jonathan B. Brown, ―Normal fasting plasma glucose and risk of Type 2 Daibetes diagnosis‖, American Journal of Medicine; 2008; 121 (6): 519-524.
[10]. 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.