Computer Architecture & Organisation (BCSE205L)

Solving Socio-Economic Problem with justification on Processor, Memory, IO and other auxiliary components

DA 1: Literature Survey - Comparative Analysis - Review Paper

NAME: R HEMESH **REG NO:22BCT0328**

TOPIC: DETECTION OF DIABETES MELLITUS USING ML

1.DOMAIN:

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood sugar levels, resulting from either insufficient insulin production or ineffective use of insulin by the body. It is a major global health concern, with an increasing prevalence over the past few decades. According to the International Diabetes Federation, approximately 463 million adults (20-79 years) were living with diabetes in 2019, and this number is expected to rise to 700 million by 2045.

Diabetes is associated with a range of complications, including cardiovascular disease, kidney failure, blindness, and lower limb amputations. Early detection and effective management of diabetes are crucial to prevent these complications and improve quality of life for patients.

Current methods for diagnosing diabetes rely on blood tests that measure fasting blood sugar levels or glucose tolerance. However, these tests may not always detect diabetes at an early stage or distinguish between different types of diabetes.

Machine learning (ML) algorithms offer a promising approach to improve the early prediction and management of diabetes. By analyzing large datasets of patient information, ML models can identify patterns and trends that may not be apparent to human experts. This can help in predicting the risk of developing diabetes, personalizing treatment plans, and improving overall patient outcomes.

In this literature survey, we will explore the application of ML algorithms, including random forest, k-nearest neighbors, logistic regression, decision tree, support vector classifier, and naive Bayes, in the prediction of diabetes mellitus. We will review existing research studies, case studies, and industry developments to understand the current state of the art, challenges, and future directions in this field.

2. LIST OF ISSUES:

1.**Early Detection:** Identifying diabetes at an early stage is crucial for effective management.

2.**Limited access to healthcare:** Not everyone has access to regular medical checkups, leading to delayed diagnoses.

3.**Feature Selection:** Choosing the most relevant features from a large dataset is crucial for developing accurate prediction models and reducing computational complexity.

4.**Class Imbalance:** Addressing the imbalance between diabetic and non-diabetic samples is essential for ensuring that prediction models are not biased towards the majority class.

5.**Interpretability:** Ensuring that ML models provide interpretable results for clinical decision-making. Understanding how machine learning models arrive at predictions is crucial for trust and clinical adoption.

6.**Generalization:** Ensuring models perform well on unseen data. Models trained on one population may not perform well on others.

7.**Data Quality:** Handling missing values, outliers, and noisy data. Inconsistent data collection and incomplete medical records can hinder prediction models.

8.**Algorithm Selection:** Choosing the most suitable machine learning algorithm for a given task is crucial for achieving optimal performance.

9.**Personalized Medicine:** Tailoring predictions and treatments to individual patient characteristics can improve outcomes and reduce healthcare costs.

10.**Ethical Considerations:** Ensuring fairness and avoiding bias in predictions. Datasets with inherent biases can lead to models that unfairly discriminate against certain demographics.

11.**Scalability:** Deploying prediction models in real-world healthcare settings requires them to be scalable and easily integrated into existing systems.

12.**Accuracy of traditional methods:** Traditional risk assessment methods may be subjective and miss prediabetic individuals, highlighting the need for more accurate and objective approaches.

13.**Data privacy & security:** Protecting sensitive patient data while using it for research and development is a challenge.

14.**Integration with clinical workflows:** Seamless integration of prediction models into healthcare systems is necessary for real-world use.

15.**Cost-effectiveness:** Implementing and maintaining machine learning solutions should be cost-effective for healthcare providers to ensure their widespread adoption and sustainability.

3. ISSUE ON FOCUS:

The motivation behind focusing on the issue of feature selection in the context of predicting diabetes mellitus using machine learning algorithms lies in the need to develop accurate and efficient prediction models. Feature selection plays a crucial role in the performance of machine learning algorithms, as it involves choosing the most relevant subset of features from a larger set of variables.

In the case of diabetes prediction, feature selection is essential for identifying the key risk factors and biomarkers associated with the disease. By selecting the most informative features, prediction models can achieve higher accuracy, better generalization to unseen data, and improved interpretability.

Moreover, efficient feature selection can help reduce the computational complexity of prediction models, making them more practical for deployment in real-world healthcare settings. This can lead to faster and more cost-effective screening and diagnostic processes, ultimately improving patient outcomes and reducing healthcare costs associated with diabetes management.

Overall, the motivation behind focusing on feature selection is to enhance the effectiveness and efficiency of machine learning algorithms in predicting diabetes mellitus, ultimately benefiting both healthcare providers and patients.

4. STATISTICS:

a. Consequences of not solving the issue - Importance and need of the hour to solve the issue:

Global health crisis: Diabetes is a major global health concern with rising prevalence. According to the World Health Organization (WHO), the number of people with diabetes has nearly quadrupled since 1980, with over 422 million adults currently affected .

Economic Burden: Diabetes imposes a significant economic burden on healthcare systems globally. The American Diabetes Association estimates the total cost of diagnosed diabetes in the US in 2023 to be $465 billion . According to the International Diabetes Federation (IDF), the global healthcare expenditure on diabetes was estimated at USD 760 billion in 2019, projected to rise to USD 825 billion by 2030. Failure to address diabetes early can lead to increased healthcare costs due to the management of complications.

Healthcare System Strain: Diabetes-related complications, such as cardiovascular disease, neuropathy, and kidney failure, can strain healthcare systems, leading to longer hospital stays, increased outpatient visits, and higher healthcare costs.

Impact on Quality of Life: Untreated or poorly managed diabetes can significantly impact an individual's quality of life, leading to reduced productivity, disability, and premature death. Uncontrolled diabetes can lead to a cascade of devastating complications, including heart disease, stroke, blindness, kidney failure, and lower limb amputation. These complications significantly reduce quality of life and increase healthcare costs.

Public Health Crisis: Diabetes is considered a global public health crisis, with the number of people affected expected to rise significantly in the coming years. Failure to address diabetes prevention and management can lead to a higher prevalence of the disease and its associated complications, further exacerbating the public health impact.

b. Citations:

1. International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: 2019. Available at: https://diabetesatlas.org/
2. [World Health Organization] Global report on diabetes https://www.who.int/health-topics/diabetes
3. [Centers for Disease Control and Prevention] National Diabetes Statistics Report https://www.cdc.gov/diabetes/data/index.html
4. [American Diabetes Association] 2023 Standards of Medical Care in Diabetes https://diabetes.org/

5. LIST OF COMPANIES:

International Companies Working on Diabetes Prediction with Machine Learning:

- **Verily Life Sciences: (US)** A subsidiary of Alphabet Inc. (Google's parent company) focused on using technology to improve healthcare. They have projects related to diabetes prediction and prevention, including utilizing machine learning to analyze health data for early risk identification.

- **Babylon Health: (UK)** Offers a digital healthcare platform that uses AI-powered chatbots to conduct symptom assessments and initial diagnoses. They are exploring the potential of using this technology for early detection of diabetes.

- **Siemens Healthineers: (Germany)** Develops medical imaging and laboratory diagnostics equipment. They are exploring how AI can be used to analyze medical images for early detection of diabetic complications.

National Companies Working on Diabetes Prediction with Machine Learning:

- **Artelus:** Based in Bengaluru, Artelus builds advanced screening tools using AI for various diseases, including diabetes. Their platform allows doctors and hospitals to conduct on-the-go screenings for early detection.

- **Wellthy Therapeutics:** This Bengaluru-based startup functions as a virtual health coach powered by AI. They offer personalized guidance on nutrition, fitness, and diabetes management, potentially aiding in early detection and prevention. Notably, Cipla, a major Indian pharmaceutical company, has invested in Wellthy, indicating a growing interest in AI-based diabetes solutions.

6. CASE STUDY:

Background:

A healthcare provider in a rural community implemented a machine learning-based diabetes prediction system to enhance early detection and management of diabetes among its patient population. The system utilized electronic health records (EHRs) and demographic information to predict the risk of developing diabetes within the next five years.

Deployment:

The machine learning model can be integrated into the existing EHR system, allowing healthcare providers to access the predictions during patient consultations. When a patient's EHR is accessed, the system generated a real-time risk score for developing diabetes and displayed it alongside the patient's medical history.

Benefits:

Early Intervention: Healthcare providers could identify individuals at high risk of developing diabetes early, enabling them to recommend preventive measures and lifestyle modifications promptly.

Patient Education: The system provided personalized information to patients about their risk factors and encouraged them to make healthier lifestyle choices to reduce their risk of developing diabetes.

Improved Patient Outcomes: Early detection and lifestyle modifications led to improved patient outcomes, such as better glycemic control and reduced risk of complications.

Efficient Use of Resources: By focusing on high-risk individuals, healthcare providers could allocate resources more efficiently, targeting those who would benefit most from preventive interventions.

Issues:

Data Quality: Ensuring the accuracy and completeness of data in EHRs was crucial for the reliability of the predictions.

Integration with Workflow: Integrating the system seamlessly into the existing workflow of healthcare providers was essential for its adoption and use in clinical practice.

Patient Engagement: Encouraging patients to actively engage with the system and follow through with recommended lifestyle modifications was a challenge.

Conclusion:

The real-time deployment of a machine learning-based diabetes prediction system in a rural healthcare setting demonstrated significant benefits in terms of early detection, patient education, and improved outcomes. Addressing issues such as data quality, workflow integration, and patient engagement is essential for the successful implementation and adoption of such systems in clinical practice.

RESEARCH PAPER - 1:

A comparison of machine learning algorithms for diabetes prediction

OBJECTIVE:

The objective of the research was to design a system that can predict diabetes with high accuracy. The researchers used seven different machine learning algorithms, including Decision Tree (DT), K Nearest Neighbor (KNN), Random Forest (RF), Naive Bayes (NB), Adaboost (AB), Logistic Regression (LR), Support Vector Machine (SVM), and a Neural Network (NN) model on the Pima Indian Diabetes Dataset (PIDD) to predict diabetes. The performance of these models was evaluated based on various measures such as accuracy, precision, recall, and F-measure. The researchers also implemented a Neural Network model for diabetic prediction of PIDD with varying hidden layers and epochs.

PROPOSED METHODOLOGY:

The researchers proposed a methodology that involves using machine learning algorithms and neural network methods to predict diabetes. The methodology is divided into several steps:

1. Data Collection and Preprocessing

The researchers used the Pima Indian Diabetes (PID) dataset, which contains information about 768 patients and their corresponding nine unique attributes. The dataset was collected from the UCI Machine Learning Repository.

Data preprocessing involved several steps:

Missing Value Identification: Missing values in the dataset were identified and replaced with the corresponding mean value.

Outlier Identification and Removal: Outliers and extreme values were detected and removed from the dataset.

Feature Selection: Pearson's correlation method was used to find the most relevant attributes/features. The correlation coefficient was calculated, which correlates with the output and input attributes.

Normalization: The data was normalized to boost the algorithm's calculation speed.

2. Dataset Train and Test Method

After data cleaning and preprocessing, the dataset was split into training and testing sets. The researchers used K-fold cross-validation and 85% train/test splitting method separately to test the different machine learning model's performance.

3. Design and Implementation of Classification Model

Different machine learning classification algorithms like Naive Bayes (NB), SVM, Linear Regression (LR), Adaboost, Random Forest, K Nearest Neighbor (KNN), Decision Tree (DT), and Neural Network (NN) with different hidden layer were used and evaluated on the dataset.

4. Neural Network Model Implementation

The researchers built three different neural network models with varying levels of hidden layers. They implemented the neural network with hidden layers 1, 2, and 3 with different epochs (200, 400, 800), and the results were compared.

The researchers used Keras and TensorFlow library to create the neural network models. They used a Sequential class from Keras library. The target variable is the 'Outcome' attribute. In ANN, the optimizer is required to reduce the output error during the backpropagation method. They used SGD (Stochastic Gradient Descent) as an optimizer. The learning rate is a parameter in an optimization algorithm that controls the weight adjustment with respect to loss gradient. They used different learning rates to find an effective one.

PERFORMANCE METRICS – MEASURE:

The performance of the machine learning algorithms used in the research was measured using several metrics derived from the confusion matrix. These metrics include:

1. **Accuracy**: This is the proportion of true results (both true positives and true negatives) in the total dataset. It is calculated using the formula:

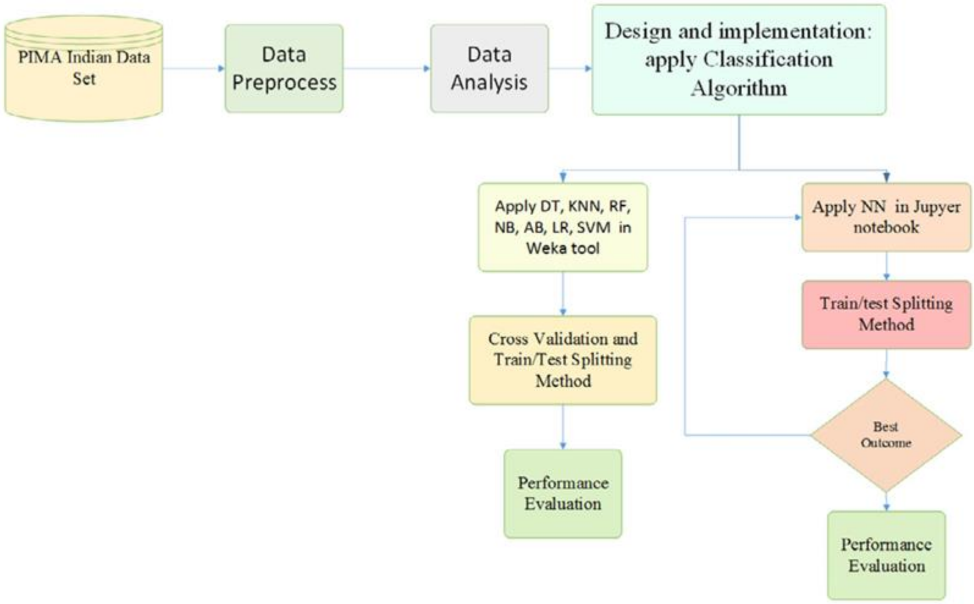
2. Accuracy = (TP + TN) / (TP + TN + FN + FP)
3. **Recall**: Also known as sensitivity, recall is the proportion of actual positives that are correctly identified. It is calculated using the formula:
4. Recall = TP / (TP + FN)
5. **Precision**: This is the proportion of positive identifications that are actually correct. It is calculated using the formula:
6. Precision = TP / (TP + FP)
7. **F-measure**: Also known as the F1 score, this is the harmonic mean of precision and recall, and it tries to find the balance between precision and recall. It is calculated using the formula:
8. F-measure = 2 * (Precision * Recall) / (Precision + Recall)

In these formulas, TP stands for True Positive, TN for True Negative, FP for False Positive, and FN for False Negative. These values are derived from the confusion matrix of the model's predictions.

ALGORITHMS:

The machine learning algorithms used include:

1. **Decision Tree (DT)**: This model predicts the value of a target variable by learning simple decision rules inferred from the data features.
2. **K-Nearest Neighbors (KNN)**: A non-parametric method used for classification and regression. The input consists of the k closest training examples in the feature space.
3. **Random Forest (RF)**: A meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve predictive accuracy and control over-fitting.
4. **Naive Bayes (NB)**: A classification technique based on Bayes’ Theorem with an assumption of independence among predictors.
5. **Adaptive Boosting (AB)**: A machine learning meta-algorithm formulated by Yoav Freund and Robert Schapire.
6. **Logistic Regression (LR)**: A statistical model that uses a logistic function to model a binary dependent variable.
7. **Support Vector Machine (SVM)**: A set of supervised learning methods used for classification, regression and outliers detection.


PIMA Indian Data Set
Data Preprocess
Data Analysis
Design and implementation: apply Classification Algorithm
Apply DT, KNN, RF, NB, AB, LR, SVM in Weka tool
Cross Validation and Train/Test Splitting Method
Performance Evaluation
Apply NN in Jupyer notebook
Train/test Splitting Method
Best Outcome
Performance Evaluation

In addition to these, the study also employed neural network methods with different hidden layers and various epochs. The researchers found that the neural network model with two hidden layers and 400 epochs provided the best accuracy, achieving an accuracy rate of 88.6%.

TEST BED:

The test bed for this research was the Pima Indian Diabetes (PID) dataset, collected from the UCI Machine Learning Repository. This dataset contains information about 768 patients and their corresponding nine unique attributes. The researchers used seven machine learning algorithms on this dataset to predict diabetes. The attributes used for the prediction of diabetes are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome. The 'outcome' attribute is taken as a dependent or target variable, and the remaining eight attributes are taken as independent/feature variables. The diabetes attribute 'outcome' consists of binary value where 0 means non-diabetes, and 1 implies diabetes.

EXPERIMENT:

The experiment conducted in the research involved using machine learning algorithms and neural network methods to predict diabetes. The researchers used the Pima Indian Diabetes (PID) dataset, which contains information about 768 patients and their corresponding nine unique attributes.

Data Preprocessing

Before the experiment, the data was preprocessed using the WEKA tool. This involved identifying and replacing missing values, identifying and removing outliers, selecting relevant features using Pearson's correlation method, and normalizing the data.

Machine Learning Algorithms

Seven machine learning algorithms were used on the dataset to predict diabetes:

1. Decision Tree (DT)
2. K Nearest Neighbor (KNN)
3. Random Forest (RF)
4. Naive Bayes (NB)
5. Adaboost (AB)
6. Logistic Regression (LR)
7. Support Vector Machine (SVM)

The performance of these algorithms was evaluated using various measures such as accuracy, precision, recall, and F-measure.

Neural Network Models

In addition to the machine learning algorithms, three different neural network models were implemented with varying levels of hidden layers (1, 2, and 3). The performance of these models was evaluated by changing the number of epochs (200, 400, 800) and learning rate (0.1, 0.01, 0.005).

RESULTS:

The results showed that all models provided an accuracy greater than 70%. The LR and SVM methods provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method. The neural network model with two hidden layers and 400 epochs provided the highest accuracy of 88.6%.

The research paper presents the results of a comparison of machine learning algorithms for diabetes prediction. The researchers used seven machine learning algorithms on the Pima Indian Diabetes (PID) dataset, which contains information about 768 patients and their corresponding nine unique attributes.

The results of the study are as follows:

- All models provided an accuracy greater than 70%.
- Logistic Regression (LR) and Support Vector Machine (SVM) provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method.

- The Neural Network (NN) model with two hidden layers provided the highest accuracy of 88.6% among all the implemented models for the PID dataset.
- The accuracy found for logistic regression (78.8571%), Naive Bayes (78.2857%), random forest (77.3429%), and ANN (88.57)% was better than the accuracy of the studies by other researchers.

In conclusion, the researchers found that the NN model with two hidden layers is the most efficient and promising for analyzing diabetes with an accuracy rate of approximately 86% for all varying epochs (200, 400, 800).

INTERPRETATION OF DATA:

The authors have used various machine learning algorithms and neural networks to predict diabetes based on the Pima Indian Diabetes dataset. The dataset contains nine unique attributes of 768 patients. After preprocessing the data, five input features (Glucose, BMI, Insulin, Pregnancy, and Age) and one output feature (outcome) were used.

The authors found that the model with Logistic Regression (LR) and Support Vector Machine (SVM) works well on diabetes prediction. They also built a Neural Network (NN) model with different hidden layers and observed that the NN with two hidden layers provided 88.6% accuracy.

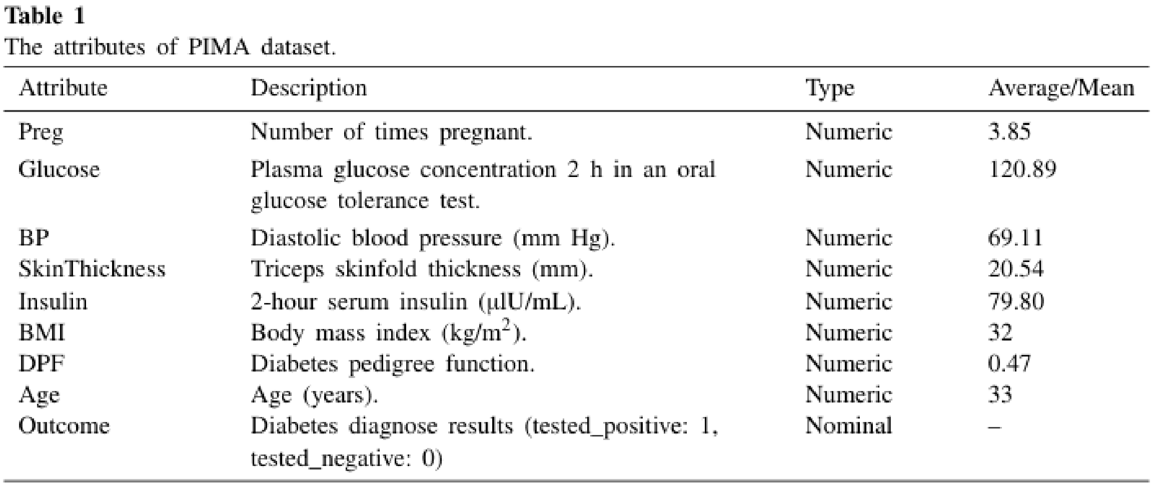


Table 1
The attributes of PIMA dataset.

Attribute	Description	Type	Average/Mean
Preg	Number of times pregnant.	Numeric	3.85
Glucose	Plasma glucose concentration 2 h in an oral glucose tolerance test.	Numeric	120.89
BP	Diastolic blood pressure (mm Hg).	Numeric	69.11
SkinThickness	Triceps skinfold thickness (mm).	Numeric	20.54
Insulin	2-hour serum insulin (μIU/mL).	Numeric	79.80
BMI	Body mass index (kg/m^2).	Numeric	32
DPF	Diabetes pedigree function.	Numeric	0.47
Age	Age (years).	Numeric	33
Outcome	Diabetes diagnose results (tested_positive: 1, tested_negative: 0)	Nominal	–

The number of missing values in PIMA dataset.

Attributes	No. of missing values
Preg	0
Glucose	5
BP	35
SkinThickness	227
Insulin	374
BMI	11
DPF	0
Age	0

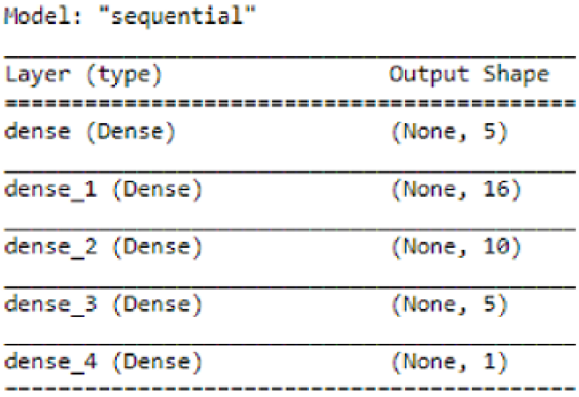
Mean and standard deviation after normalization.

Impact of learning rate on accuracy measurement.

Learning rate	Accuracy
0.1	0.829
0.01	**0.838**
0.005	0.800

At learning rate 0.01 with hidden layer changes impact on the accuracy.

Hidden layer	Epochs	Accuracy	Training accuracy	Testing accuracy
1	200	0.838	76.43%	83.81%
	400	0.848	77.27%	84.76%
	800	0.829	79.46%	82.86%
2	200	0.876	76.77%	87.62%
	400	**0.886**	78.96%	**88.57%**
	800	0.857	81.65%	87.62%
3	200	0.829	76.77%	82.86%
	400	0.838	83.00%	83.81%
	800	0.790	87.04%	79.05%



```
Model: "sequential"
_____________________________________________
Layer (type)                 Output Shape
=============================================
dense (Dense)                (None, 5)
_____________________________________________
dense_1 (Dense)              (None, 16)
_____________________________________________
dense_2 (Dense)              (None, 10)
_____________________________________________
dense_3 (Dense)              (None, 5)
_____________________________________________
dense_4 (Dense)              (None, 1)
---------------------------------------------
```

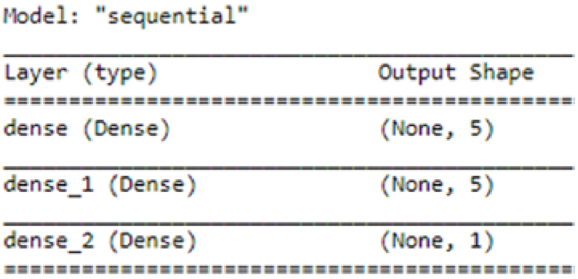
NN model with three hidden layers.

The correlation between input and output attributes.

Attributes	Correlation coefficient
Glucose	0.484
BMI	0.316
Insulin	0.261
Preg	0.226
Age	0.224
SkinThickness	0.193
BP	0.183
DPF	0.178

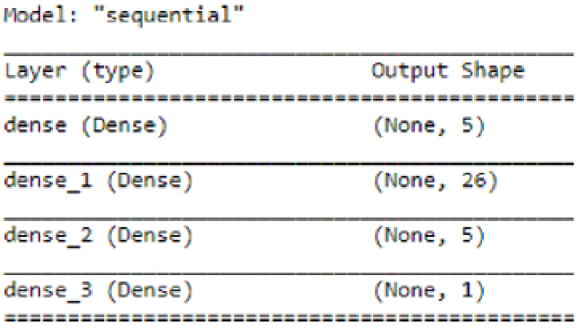
The performance measure of all classification methods for K-fold cross-validation and Train/Test splitting method.

Classification	Precision	Recall	F-measure	Accuracy
DT (K-fold)	0.739	0.742	0.741	74.24%
DT (Splitting)	0.735	0.731	0.733	73.14%
RF (K-fold)	0.744	0.750	0.746	74.96%
RF (Splitting)	0.779	0.771	0.774	77.14%
NB (K-fold)	0.753	0.755	0.754	75.53%
NB (Splitting)	0.787	0.783	0.785	78.28%
LR (K-fold)	0.761	0.768	0.761	76.82%
LR (Splitting)	0.788	0.789	0.788	78.85%
KNN (K-fold)	0.747	0.751	0.749	75.10%
KNN (Splitting)	0.804	0.794	0.798	79.42%
AB (K-fold)	0.730	0.740	0.730	73.96%
AB (Splitting)	0.792	0.794	0.793	79.42%
SVM (K-fold)	0.761	0.768	0.759	76.82%
SVM (Splitting)	0.774	0.777	0.775	77.71%



```
Model: "sequential"
_____________________________________________
Layer (type)                 Output Shape
=============================================
dense (Dense)                (None, 5)
_____________________________________________
dense_1 (Dense)              (None, 5)
_____________________________________________
dense_2 (Dense)              (None, 1)
=============================================
```

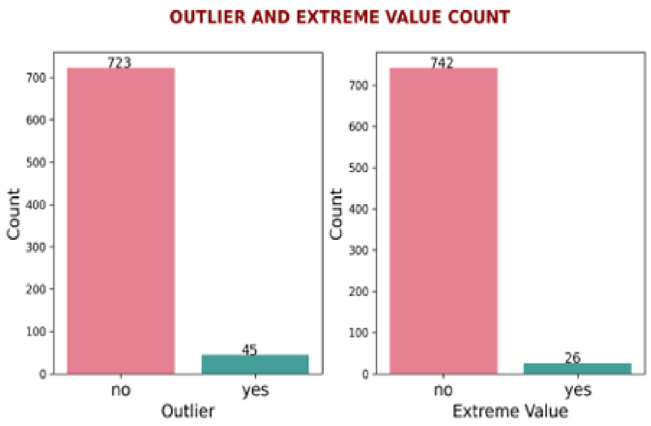
NN model with one hidden layer.



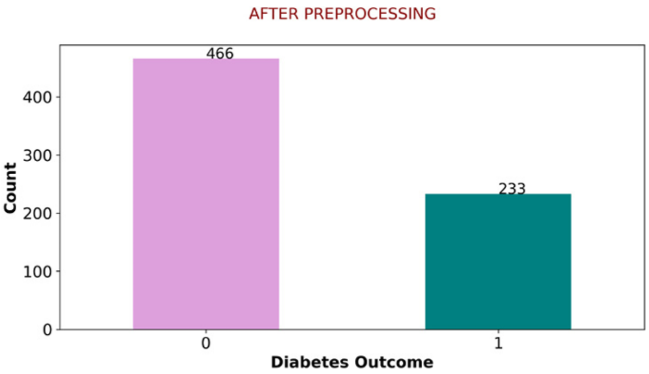
```
Model: "sequential"
_____________________________________________
Layer (type)                 Output Shape
=============================================
dense (Dense)                (None, 5)
_____________________________________________
dense_1 (Dense)              (None, 26)
_____________________________________________
dense_2 (Dense)              (None, 5)
_____________________________________________
dense_3 (Dense)              (None, 1)
=============================================
```

NN model with two hidden layers.

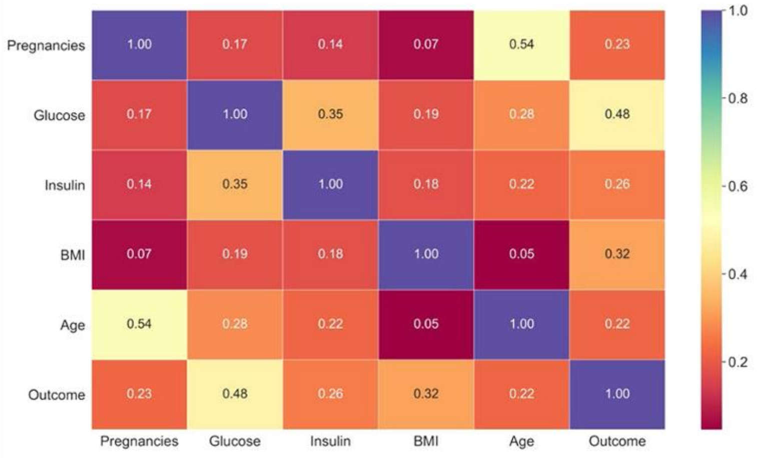
INTERPRETATION OF GRAPHS:



Using the Weka tool, the data set for detecting outliers and extreme values based on inter quartile ranges is filtered.The number of outliers and extreme values are shown in figure, where we can see that there are 45 outliers and 26 extreme values.

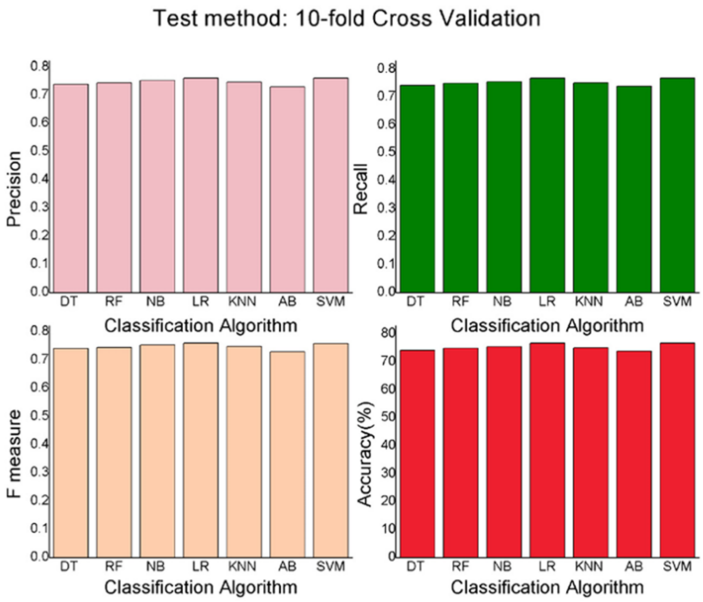


After Preprocessing the Number of Diabetes and Non-Diabetes Patients : This bar graph shows the number of patients with and without diabetes after preprocessing the data. It shows that there are more non-diabetic patients than diabetic patients in the dataset.



After preprocessing correlation between input and output attributes.

After Preprocessing Correlation Between Input and Output Attributes : This shows the correlation between the input features and the output feature (diabetes outcome). The 'Glucose' feature has the highest correlation with the outcome, indicating that it is a significant factor in predicting diabetes.

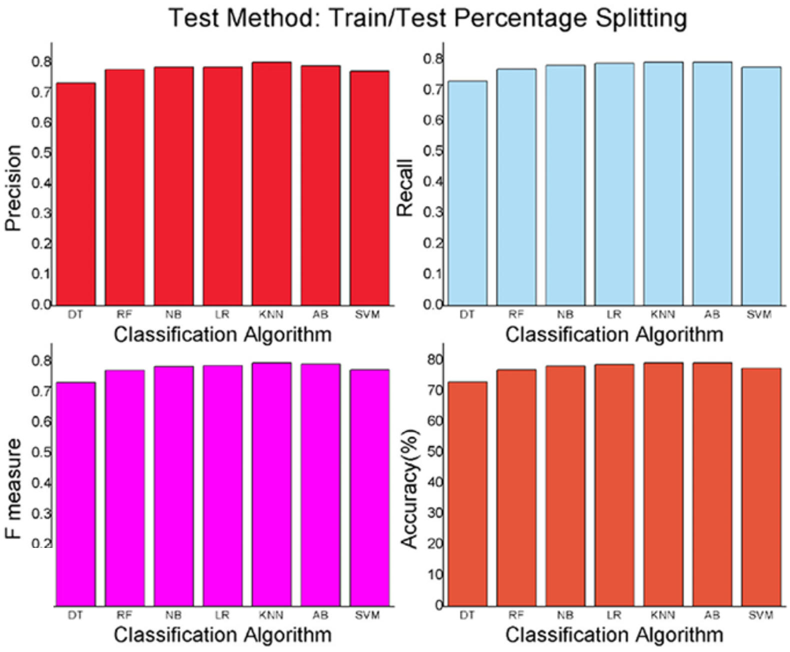


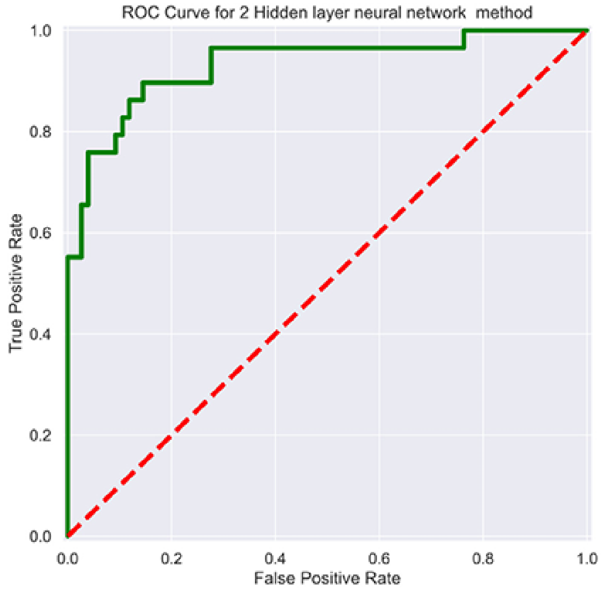
Graphical Presentation of the Performance of All Classifiers with a 10-fold Cross-Validation Method

This line graph shows the performance of all classifiers using a 10-fold cross-validation method. It shows that the Logistic Regression (LR) and Support Vector Machine (SVM) classifiers have the highest accuracy.

Graphical Presentation of the Performance of Classifier with Train/Test Splitting Method :

This line graph shows the performance of all classifiers using a train/test splitting method. It shows that the K-Nearest Neighbors (KNN) and AdaBoost (AB) classifiers have the highest accuracy.





ROC curve for 2 hidden layer NN with 400 epochs.

ROC Curve for 2 Hidden Layer NN with 400 Epochs :

This graph shows the Receiver Operating Characteristic (ROC) curve for the Neural Network model with two hidden layers and 400 epochs. The area under the curve is large, indicating that the model has a high true positive rate and a low false positive rate, which means it has good performance.

PROS:

1. **Early Detection of Diabetes**: The research paper highlights the potential of machine learning algorithms and neural networks in early detection of diabetes. This is a significant advantage as early detection is key to effective management and treatment of the disease.
2. **High Accuracy**: The research paper reports that the Neural Network model with two hidden layers provided the highest accuracy (88.6%) among all the implemented models for the Pima Indian Diabetes dataset. This high accuracy rate is a strong point in favor of the methods used in the study.
3. **Use of Multiple Algorithms**: The study's comprehensive analysis, which involved the use of seven different machine learning algorithms, is another positive aspect. This approach allows for a more robust and reliable evaluation of performance.
4. **Data Preprocessing**: The research paper also emphasizes the importance of data preprocessing techniques like outlier identification, missing value identification, feature selection, and normalization. These techniques can significantly improve the quality of data, leading to more accurate results.

CONS:

1. **Complexity**: On the downside, implementing machine learning algorithms and neural networks requires a deep understanding of these techniques. This complexity can pose a challenge, especially for those who are new to these methods.
2. **Data Sensitivity**: The research paper also points out that the accuracy of predictions heavily depends on the quality of the data. Any errors or inconsistencies in the data can lead to inaccurate predictions, which is a significant drawback.
3. **Computational Resources**: Another disadvantage is that machine learning algorithms, especially neural networks, can be computationally intensive. This means they require significant processing power and memory, which may not always be readily available.
4. **Overfitting Risk**: Lastly, there is a risk of overfitting with machine learning models. This means that the model may perform well on the training data but poorly on new, unseen data, which can limit its practical applicability.

FUTURE WORKS:

The future works of this research paper could include the following:

1. **Feature Engineering**: Exploring additional features or engineering new features that could enhance the predictive power of the models. This could involve domain-specific knowledge to identify relevant features related to diabetes prediction.
2. **Ensemble Methods**: Investigating the use of ensemble methods, such as stacking or boosting, to combine the predictions of multiple models for improved accuracy.
3. **Hyperparameter Tuning**: Conducting a more extensive search for optimal hyperparameters for each algorithm to further improve their performance.
4. **Imbalanced Data Handling**: Addressing the issue of imbalanced data by exploring techniques such as oversampling, undersampling, or using different evaluation metrics to account for class imbalances.
5. **Deep Learning**: Exploring the use of deep learning models, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), for diabetes prediction, especially if there are large amounts of data available.
6. **External Validation**: Validating the models on external datasets to assess their generalizability and robustness across different populations and data sources.
7. **Clinical Implementation**: Investigating the practical implementation of the predictive models in clinical settings, including considerations for interpretability, explainability, and integration with existing healthcare systems.
8. **Ethical Considerations**: Addressing ethical considerations related to the use of predictive models in healthcare, including privacy, transparency, and fairness in model predictions.
9. **Longitudinal Analysis**: Conducting longitudinal studies to assess the predictive power of the models over time and to understand the progression of diabetes in patients.
10. **Interpretability**: Exploring methods to enhance the interpretability of the models, especially for complex models like neural networks, to provide actionable insights for healthcare professionals.

These future works aim to advance the research in diabetes prediction by addressing specific challenges, improving model performance, and facilitating the practical application of predictive models in clinical settings.

REFERENCES:

1. World Health Organization. (n.d.). Diabetes. Retrieved from https://www.who.int/health-topics/diabetes

2. Medical News Today. (n.d.). How is the pancreas linked with diabetes? Retrieved from https://www.medicalnewstoday.com/articles/325018#how-is-the-pancreas-linked-with-diabetes

3. WebMD. (n.d.). Diabetes Causes. Retrieved from https://www.webmd.com/diabetes/diabetes-causes

4. Mayo Clinic. (n.d.). Diagnosis & treatment of prediabetes. Retrieved from https://www.mayoclinic.org/diseases-conditions/prediabetes/diagnosis-treatment/drc-20355284

5. National Institute of Diabetes and Digestive and Kidney Diseases. (n.d.). Symptoms & Causes of Diabetes. Retrieved from https://www.niddk.nih.gov/health-information/diabetes/overview/symptoms-causes

6. Diabetes.co.uk. (n.d.). Blood Sugar Level Ranges. Retrieved from https://www.diabetes.co.uk/diabetes_care/blood-sugar-level-ranges.html

7. Healthgrades. (n.d.). Is There a Cure for Diabetes? Retrieved from https://www.healthgrades.com/right-care/diabetes/is-there-a-cure-for-diabetes

8. Better Health Channel. (n.d.). Diabetes - long-term effects. Retrieved from https://www.betterhealth.vic.gov.au/health/conditionsandtreatments/diabetes-long-term-effects

9. Centers for Disease Control and Prevention. (n.d.). National Diabetes Prevention Program. Retrieved from https://www.cdc.gov/diabetes/basics/prediabetes.html

10. Lichman, M. (n.d.). Pima Indians Diabetes Database. UCI Machine Learning Repository. Retrieved from https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

11. Wikipedia. (n.d.). Project Jupyter. Retrieved from https://en.wikipedia.org/wiki/Project_Jupyter

PUBLISHED PAPER:

https://www.sciencedirect.com/science/article/pii/S2405959521000205

RESEARCH PAPER - 2:

A Neural Network based Diabetes Prediction on Imbalanced Data

OBJECTIVE:

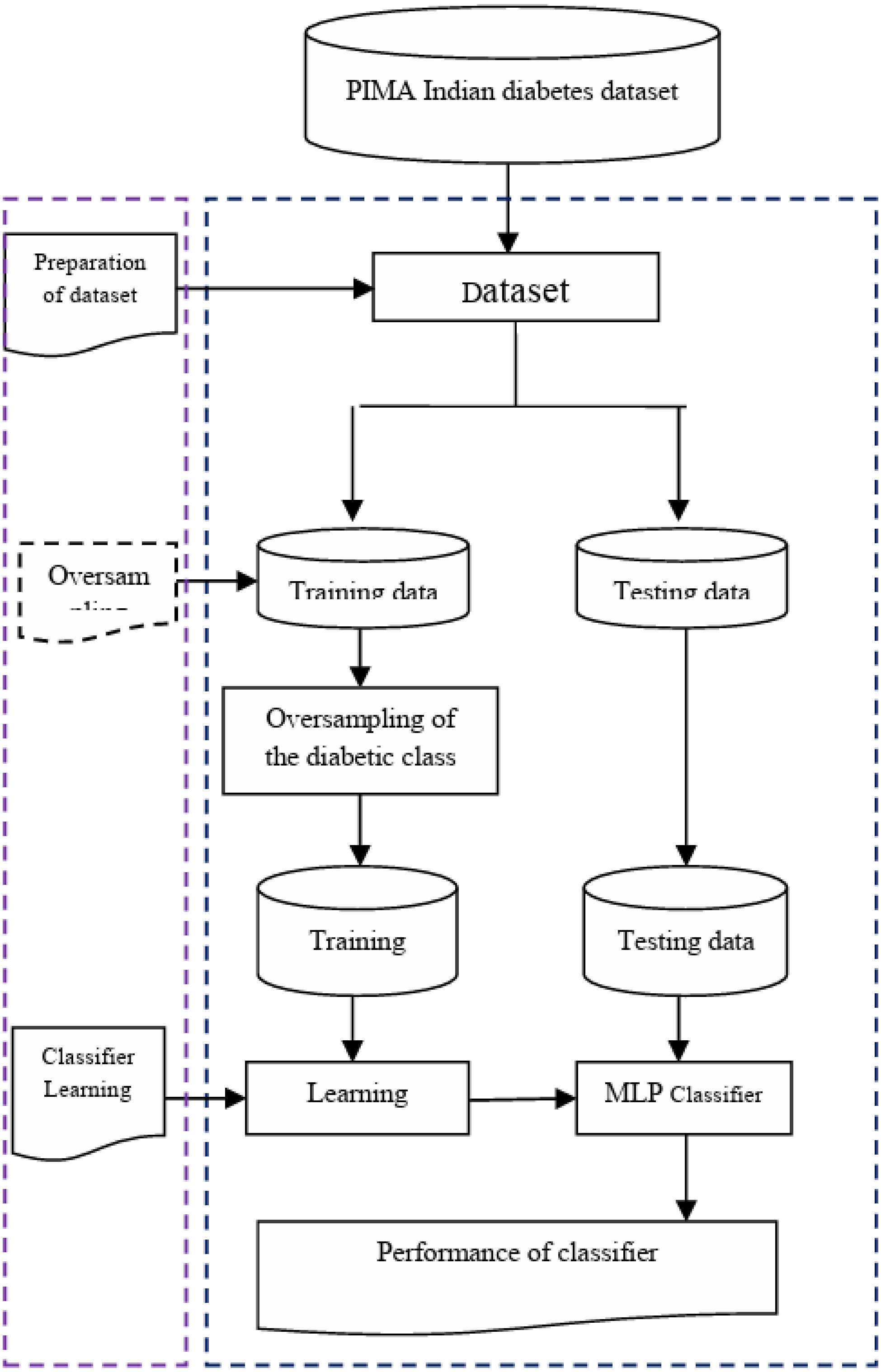
The objective of the research paper is to develop a machine learning (ML) algorithm for the early detection and prediction of diabetes using the PIMA Indian diabetes dataset (PIDD) obtained from the University of California, Irvine (UCI) ML repository. The focus is on addressing the challenges posed by imbalanced and missing data in the dataset, as well as the presence of outliers. The goal is to improve the performance of the prediction model for diabetes classification by implementing preprocessing techniques and a multilayer perceptron (MLP) classifier.

PROPOSED METHODOLOGY:

The proposed methodology in the text is a machine learning model for predicting diabetes. The model is designed to address several challenges in diabetes prediction, including the presence of missing values, outliers, and imbalanced class distribution in medical records. Here are the key steps in the proposed methodology:

1. **Data Preparation:** The model uses the PIMA Indian Diabetes Dataset, which is split into training and testing data.
2. **Data Preprocessing:** This step involves several processes:
   - Outlier Detection and Replacement: Outliers in the data are detected using the Interquartile Range (IQR) method and replaced with the median value of the attributes.
   - Missing Value Imputation: Missing values in the data are filled using the mean value of the attributes.
   - Feature Scaling: This process normalizes the data within a range.
   - Feature Selection: The one-way ANOVA F-test is used to reduce the high dimensionality of the feature space before the classification process.
3. **Oversampling:** The Adaptive Synthetic (ADASYN) oversampling method is used to integrate strongly similar data points through K-Nearest Neighbors (KNN).

4. **Classification:** The Multilayer Perceptron (MLP) classifier is used for predicting the classification of diabetes with the K-fold cross-validation technique.
5. **Evaluation:** The model's performance is evaluated using several metrics, including accuracy, precision, recall, F-measures, and the area under the curve (AUC).



The proposed methodology aims to improve the accuracy of diabetes prediction by addressing the challenges of missing values, outliers, and imbalanced class distribution in the dataset.

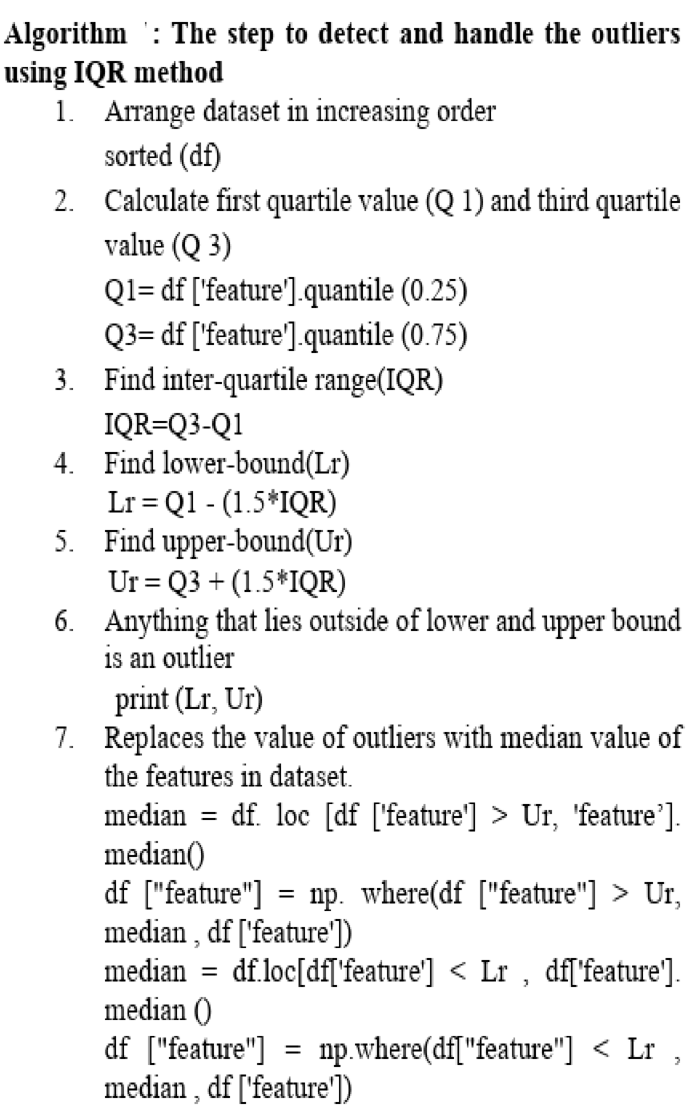
PERFORMANCE METRICS – MEASURE:

The performance of the proposed model in the given text is evaluated using several metrics. These metrics are calculated based on the confusion matrix, which consists of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. The metrics used are:

1. **Accuracy**: This is the ratio of the number of correctly classified diabetes patients to the total number of diabetes datasets. It is calculated using the formula:
2. Accuracy = (TN + TP) / (TP + TN + FP + FN)
3. **Precision**: This is the ratio of the number of positively classified diabetic patients to the number of samples predicted as having diabetes. It is calculated using the formula:
4. Precision = TP / (TP + FP)
5. **Recall**: This is the ratio of truly positive diabetes datasets to the number of actual positive diabetic patients. It is calculated using the formula:
6. Recall = TP / (TP + FN)
7. **F-measure**: This is the harmonic mean of accuracy and precision, which works effectively for diagnosing tests. It is calculated using the formula:
8. F-measure = (2 * Precision * Recall) / (Precision + Recall)
9. **Area Under the Curve (AUC)**: This is used for reporting a well-ranked prediction instead of absolute values. It shows the performance of the model at various threshold settings.

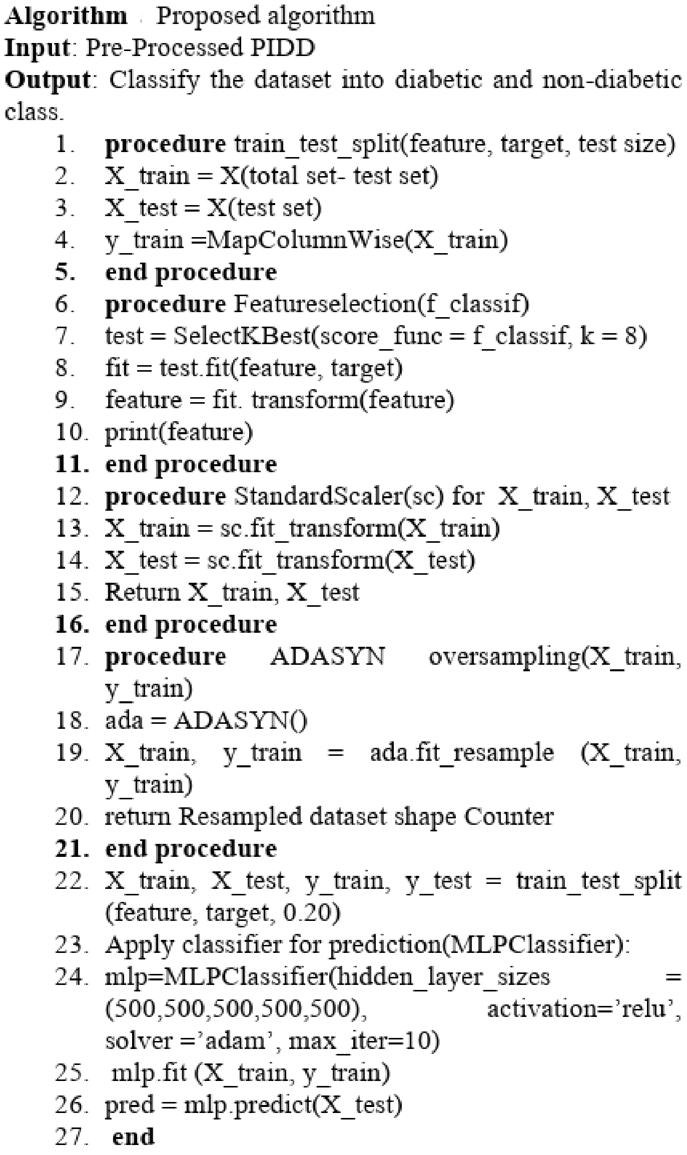
These metrics provide a comprehensive evaluation of the model's performance, taking into account both the model's accuracy and its ability to correctly classify positive cases.

ALGORITHMS:



Algorithm ' : The step to detect and handle the outliers using IQR method

1. Arrange dataset in increasing order
   sorted (df)
2. Calculate first quartile value (Q 1) and third quartile value (Q 3)
   Q1= df ['feature'].quantile (0.25)
   Q3= df ['feature'].quantile (0.75)
3. Find inter-quartile range(IQR)
   IQR=Q3-Q1
4. Find lower-bound(Lr)
   Lr = Q1 - (1.5*IQR)
5. Find upper-bound(Ur)
   Ur = Q3 + (1.5*IQR)
6. Anything that lies outside of lower and upper bound is an outlier
   print (Lr, Ur)
7. Replaces the value of outliers with median value of the features in dataset.
   median = df. loc [df ['feature'] > Ur, 'feature']. median()
   df ["feature"] = np. where(df ["feature"] > Ur, median , df ['feature'])
   median = df.loc[df['feature'] < Lr , df['feature']. median ()
   df ["feature"] = np.where(df["feature"] < Lr , median , df ['feature'])



Algorithm . Proposed algorithm
Input: Pre-Processed PIDD
Output: Classify the dataset into diabetic and non-diabetic class.

```
procedure train_test_split(feature, target, test size)
X_train = X(total set- test set)
X_test = X(test set)
y_train =MapColumnWise(X_train)
end procedure
procedure Featureselection(f_classif)
test = SelectKBest(score_func = f_classif, k = 8)
fit = test.fit(feature, target)
feature = fit. transform(feature)
print(feature)
end procedure
procedure StandardScaler(sc) for  X_train, X_test
X_train = sc.fit_transform(X_train)
X_test = sc.fit_transform(X_test)
Return X_train, X_test
end procedure
procedure ADASYN oversampling(X_train, y_train)
ada = ADASYN()
X_train, y_train = ada.fit_resample (X_train, y_train)
return Resampled dataset shape Counter
end procedure
X_train, X_test, y_train, y_test = train_test_split (feature, target, 0.20)
Apply classifier for prediction(MLPClassifier):
mlp=MLPClassifier(hidden_layer_sizes = (500,500,500,500,500), activation='relu', solver ='adam', max_iter=10)
 mlp.fit (X_train, y_train)
pred = mlp.predict(X_test)
 end
```

1. Interquartile Range (IQR) Method

This algorithm is used to detect and handle outliers in the data. It involves arranging the data in increasing order, calculating the first and third quartiles (Q1 and Q3), and then finding the interquartile range (IQR). The lower and upper bounds are then calculated, and any data outside these bounds is considered an outlier. Outliers are replaced with the median value of the attributes in the dataset.

2. Adaptive Synthetic Sampling (ADASYN)

This is a resampling technique used to handle imbalanced data. It generates synthetic data based on the minority class to balance the class distribution.

3. Multilayer Perceptron (MLP) Classifier

This is a type of artificial neural network used for the classification task. It consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The MLP classifier in the text uses five hidden layers and is trained using a supervised machine learning technique.

4. K-Fold Cross-Validation

This is a technique used to assess the performance of the MLP classifier. It involves dividing the dataset into 'k' subsets, and then iteratively training the model on 'k-1' subsets and testing it on the remaining subset.

5. One-Way ANOVA F-Test

This statistical method is used for feature selection. It helps to select the features that are highly related to the outcome variable.

6. Other Machine Learning Techniques

The text also mentions other machine learning techniques used in previous research for diabetes prediction, including Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB), k-Nearest Neighbor (KNN), and Random Forest (RF).

TEST BED:

The test bed for the research conducted in the provided text is the PIMA Indian Diabetes Dataset (PIDD) taken from the UCI Repository. This dataset consists of 768 instances with 8 attributes, only 113 data without missing values. It includes 268 diabetic patients and 500 non-diabetic patients, which presents a class imbalance in the dataset. The dataset includes 8 numerical value features such as pregnant count, glucose concentration, blood pressure (mm Hg), Skin Thickness (mm), insulin (mm U/ml), body mass index (BMI), Diabetes pedigree function (PDF), and Age (years). The '0' in the dataset represents the negative value of diabetes, and '1' represents a positive value of diabetes.

EXPERIMENT:

The experiment was conducted using the PIMA Indian Diabetes Dataset (PIDD) from the University of California, Irvine (UCI) Machine Learning repository. The dataset consists of 768 instances with 8 attributes, including the number of pregnancies, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The dataset also includes a class label, where '0' represents the negative value of diabetes, and '1' represents a positive value.

The experiment involved several steps:

1. **Data Preparation**: The PIDD was divided into training and testing datasets.
2. **Outlier Detection and Replacement**: Outliers in the data were detected using the Interquartile Range (IQR) method and replaced with the median value of the attributes.
3. **Missing Value Imputation**: Missing values in the data were imputed using the mean value of the attributes.
4. **Feature Scaling and Selection**: The features of the dataset were scaled and selected using the one-way ANOVA F-test.
5. **Oversampling**: The Adaptive Synthetic Sampling (ADASYN) method was used to oversample the minority class in the dataset.
6. **Classification**: A Multilayer Perceptron (MLP) classifier was used to classify the data into diabetic or non-diabetic based on their symptoms.

The performance of the proposed model was evaluated using various metrics such as accuracy, precision, recall, F-measure, and Area Under the Curve (AUC). The results showed that the proposed model achieved an accuracy of 84%, outperforming other benchmark algorithms.

RESULTS:

The results of the experiment demonstrate the effectiveness of the proposed algorithm for diabetes prediction. The performance metrics including accuracy, precision, recall, F-measure, and AUC are used to evaluate the classifier's performance. The comparison of the proposed algorithm's results with benchmark algorithms shows that the proposed model outperforms the other algorithms, indicating its efficacy in predicting diabetes.

1. **Accuracy**: The proposed algorithm achieved an accuracy of 84%, indicating the ratio of correctly classified diabetes patients to the total number of diabetes datasets.
2. **Precision**: The precision of the proposed algorithm is 91%, which represents the ratio of positively classified diabetic patients to the number of samples predicted as diabetic.
3. **Recall**: The recall of the proposed algorithm is 85%, indicating the ratio of truly positive diabetes datasets to the number of actual positive diabetic patients.

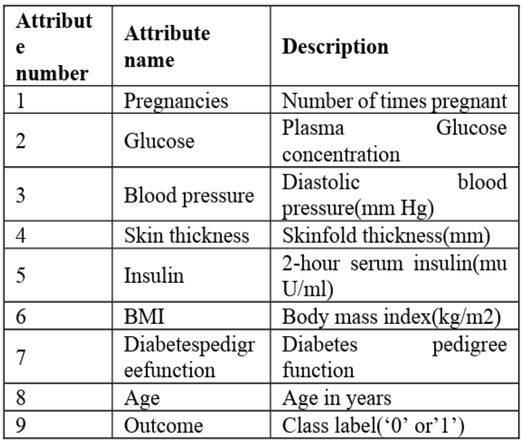
4. **F-measure**: The F-measure of the proposed algorithm is 88%, which represents the harmonic mean of accuracy and precision, providing a balanced measure of the classifier's performance.
5. **Area Under the Curve (AUC)**: The AUC score of the proposed algorithm is 83%, which is a metric used to evaluate the overall performance of the classifier, particularly in binary classification problems.

These results demonstrate that the proposed algorithm yields high accuracy, precision, recall, and F-measure, indicating its effectiveness in predicting diabetes. The AUC score further validates the robustness of the classifier in distinguishing between diabetic and non-diabetic patients.

The comparison of the proposed algorithm's results with benchmark algorithms shows that the proposed model outperforms other algorithms, indicating its superiority in predicting diabetes based on the PIMA Indian diabetes dataset.

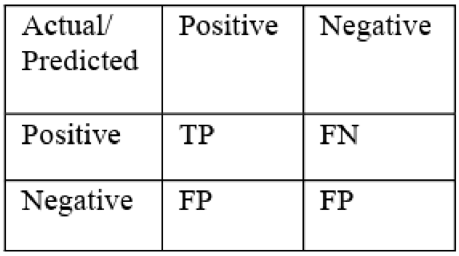
INTERPRETATION OF DATA:

Table 1 Description of PIDD



Attribute number	Attribute name	Description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma Glucose concentration
3	Blood pressure	Diastolic blood pressure(mm Hg)
4	Skin thickness	Skinfold thickness(mm)
5	Insulin	2-hour serum insulin(mu U/ml)
6	BMI	Body mass index(kg/m2)
7	Diabetespedigreefunction	Diabetes pedigree function
8	Age	Age in years
9	Outcome	Class label('0' or'1')

Table 2 Confusion Matrix



Actual/ Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	FP

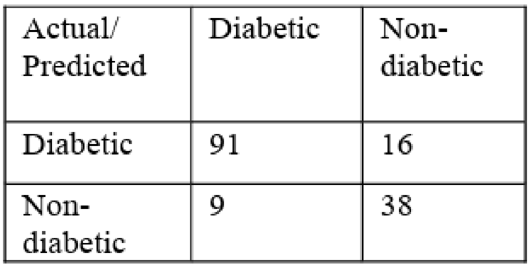
$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F - measure = \frac{(\ 2 \times Precision \times Recall\ )}{(Precision + Recall)} \tag{9}$$

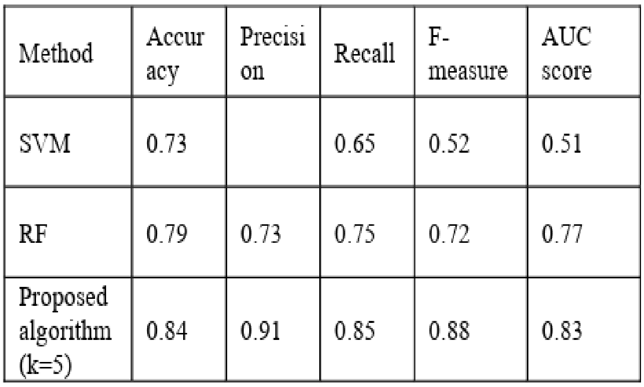
Table 3 Results in Confusion Matrix



Actual/ Predicted	Diabetic	Non-diabetic
Diabetic	91	16
Non-diabetic	9	38

This table shows the results of the confusion matrix after applying the proposed algorithm. The model correctly identified 91 diabetic cases (TP) and 38 non-diabetic cases (TN). It incorrectly identified 16 diabetic cases as non-diabetic (FN) and 9 non-diabetic cases as diabetic (FP).

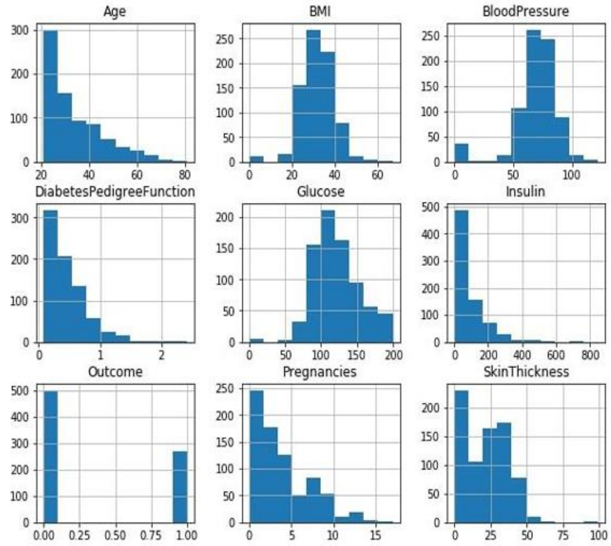
Table 4 Results Comparison



Method	Accuracy	Precision	Recall	F-measure	AUC score
SVM	0.73		0.65	0.52	0.51
RF	0.79	0.73	0.75	0.72	0.77
Proposed algorithm (k=5)	0.84	0.91	0.85	0.88	0.83

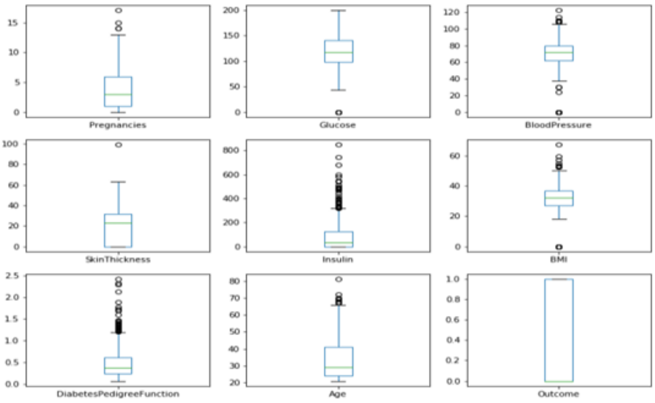
This table compares the performance of different machine learning methods (SVM, RF, and the proposed algorithm) in terms of accuracy, precision, recall, F-measure, and AUC score. The proposed algorithm outperforms the other methods in all metrics.

INTERPRETATION OF GRAPHS:



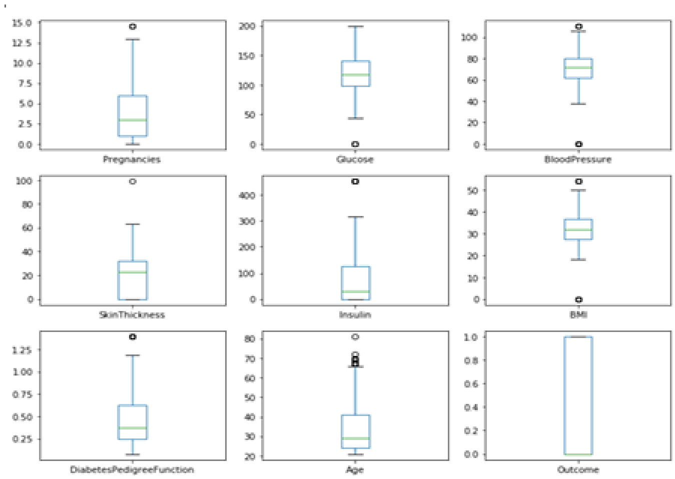
Distribution of Attributes in PIDD

These attributes include numerical values such as pregnant count, glucose concentration, blood pressure, skin thickness, insulin level, body mass index (BMI), Diabetes pedigree function (PDF), and age. The distribution is shown through various bar graphs to provide insights into the range, central tendency, and dispersion of values for each attribute. This helps in understanding the data better and in making decisions about data preprocessing steps like outlier detection, missing value imputation, and feature scaling.



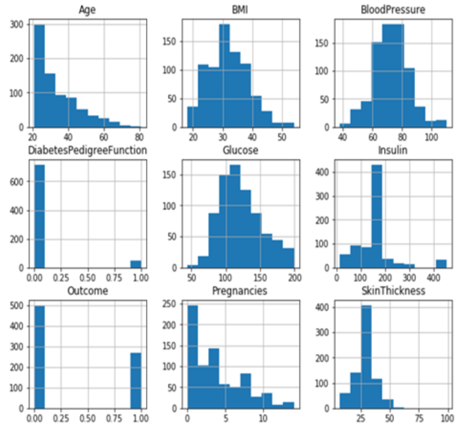
. Presence of outliers in PIDD

The figure is a boxplot, a standard way of graphically depicting groups of numerical data through their quartiles.In this study, outliers are data points that significantly deviate from other observations. These outliers can affect the performance of the machine learning model, leading to less accurate predictions. Therefore, identifying and handling these outliers is a crucial step in the data preprocessing stage of building the prediction model.



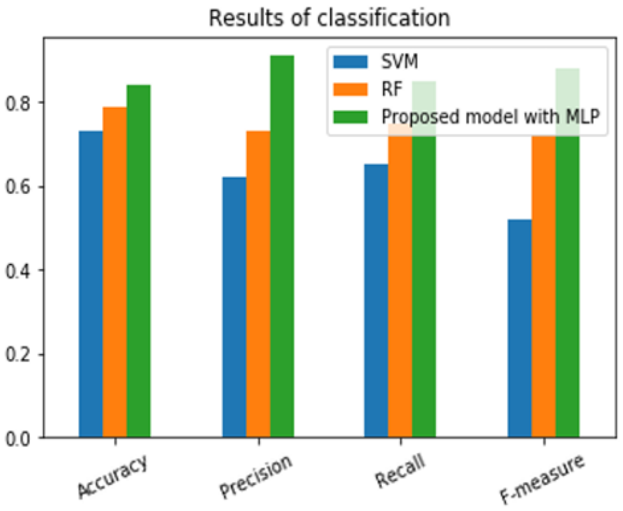
After replacement of outliers in PIDD

This is a visual representation of the dataset after the replacement of outliers. Outliers are extreme values that deviate significantly from other observations in the dataset and can affect the performance of machine learning models. In this case, the outliers in the PIMA Indian Diabetes Dataset (PIDD) were replaced with the median value of the attributes, instead of removing them. This process is part of the data preprocessing step to improve the performance of the machine learning model. The figure shows a more balanced and normalized distribution of the data after the outlier replacement.



Distribution of Attributes after Applying ADASYN

ADASYN is an oversampling technique used in machine learning to balance the dataset. It generates synthetic samples in a dataset to balance the minority class, which in this case is the diabetic class. The figure shows the distribution of eight numerical value features such as pregnant count, glucose concentration, blood pressure (mm Hg), skin thickness (mm), insulin (mm U/ml), body mass index (BMI), diabetes pedigree function (PDF), and age (years). After applying the ADASYN method, the distribution of these attributes changes, which can potentially improve the performance of the prediction model.



Graphs Represents Comparison of Results

This is a graphical representation that compares the performance of different machine learning methods used for diabetes prediction. The methods compared include Support Vector Machine (SVM), Random Forest (RF), and the proposed algorithm in the paper.

The graph shows the performance of these methods in terms of various evaluation metrics such as accuracy, precision, recall, F-measure, and Area Under the Curve (AUC) score. These metrics are used to evaluate the performance of the classification models.

From the graph, it can be inferred that the proposed algorithm outperforms the other two methods (SVM and RF) across all the evaluation metrics. This suggests that the proposed algorithm is more effective in predicting diabetes based on the PIMA Indian diabetes dataset.

PROS:

1. **Efficient Prediction**: The research paper presents a machine learning algorithm that can efficiently predict the presence of diabetes using the PIMA Indian diabetes dataset.
2. **Handling Imbalanced Data**: The algorithm effectively handles imbalanced and missing data, which are common issues in medical datasets.
3. **Improved Accuracy**: The experiment obtained a better accuracy of 84% with a neural network model in comparison with the previous model.
4. **Outlier and Missing Value Handling**: The proposed model includes steps for detecting and handling outliers and missing values, which can improve the performance of the prediction model.
5. **Feature Selection**: The model uses a feature selection technique to improve generalization and make better predictions.

CONS:

1. **Data Limitations**: The model is trained and tested on the PIMA Indian diabetes dataset. The performance of the model might vary when used with different datasets.
2. **Preprocessing Requirement**: The model requires several preprocessing steps like detection of outliers, filling missing values, features scaling, and feature selection of attributes. This could be time-consuming and computationally expensive.
3. **Dependence on Hyperparameters**: The performance of the Multilayer Perceptron (MLP) classifier used in the model depends on the optimization of hyperparameters, which can be a complex task.
4. **Class Imbalance**: The presence of class imbalance in the dataset can affect the performance of the classifier. The model uses an oversampling technique to handle this, but it might not always be the best solution for all datasets.

FUTURE WORKS:

Future work could potentially involve further refining the machine learning model for diabetes prediction, exploring other machine learning algorithms, or applying the model to other medical datasets. Additionally, the model could be improved by incorporating more features or using more advanced techniques for handling missing data and outliers.

REFERENCES:

1. M. Sued, S. Lara, L. Abdurrahman,R. Almohaini, and T. Saba. Current Techniques for Diabetes Prediction: Review and Case Study. Applied. Science. (2019) https://www.mdpi.com/2076-3417/9/15/3169
2. K. Sumengalli, Gitika, S.B.R., and H.Ambharkar. A classifier based method for the earliest detection of diabetic or non-diabetic. In: 2016 International Conference on Control, Instrumentation, Communication, and Computational Technologies (ICCICCT). IEEE (2016)

   https://ieeexplore.ieee.org/abstract/document/7987974
3. American Diabetes Association. "Diagnosis and Classification of Diabetes." Diabetes Care (May 2018) https://care.diabetesjournals.org/content/41/Supplement_1/S13
4. P.Shonar K.MaliniJaya, Diabetes Prediction Using Different ML Approaches.3rd International Conference on Computing Methodologies and Communication (ICCMC), (2019)

   https://ieeexplore.ieee.org/abstract/document/8826128
5. PIMA Indian diabetes dataset: https://archive.ics.uci.edu/ml/dataset-PIMA-indian-datasets/ 5

PUBLISHED PAPER:

https://ieeexplore-ieee-org.egateway.vit.ac.in/document/9509732

RESEARCH PAPER - 3:

Prediction of Diabetes using Classification Algorithms:

OBJECTIVE:

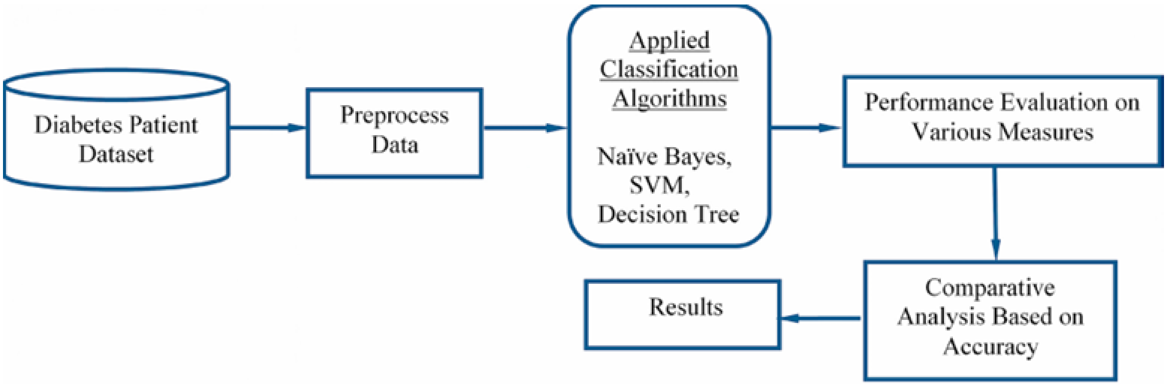
The objective of this study is to design a model that can predict the likelihood of diabetes in patients with maximum accuracy. The researchers used three machine learning classification algorithms, namely Decision Tree, SVM, and Naive Bayes, to detect diabetes at an early stage. The experiments were performed on the Pima Indians Diabetes Database (PIDD) sourced from the UCI machine learning repository. The performance of all three algorithms was evaluated based on various measures like Precision, Accuracy, F-Measure, and Recall.

PROPOSED METHODOLOGY:

The study proposed a methodology to predict the likelihood of diabetes in patients using machine learning classification algorithms. The three algorithms used in the experiment were Decision Tree, SVM (Support Vector Machine), and Naive Bayes.

Model Diagram

The proposed procedure was summarized in a model diagram, which showed the flow of the research conducted in constructing the model.


Diabetes Patient Dataset
Preprocess Data
Applied Classification Algorithms
Naïve Bayes, SVM, Decision Tree
Performance Evaluation on Various Measures
Comparative Analysis Based on Accuracy
Results

Brief Description of Algorithms Used

Support Vector Machine (SVM)

SVM is a supervised machine learning model used in classification. It aims to find the best highest-margin separating hyperplane between two classes. The hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors.

Naive Bayes Classifier

Naive Bayes is a classification technique that assumes all features are independent and unrelated to each other. It works well for data with imbalancing problems and missing values. Naive Bayes is a machine learning classifier which employs the Bayes Theorem.

Decision Tree Classifier

Decision Tree is a supervised machine learning algorithm used to solve classification problems. It uses nodes and internodes for prediction and classification. Root nodes classify the instances with different features. In every stage, the Decision tree chooses each node by evaluating the highest information gain among all the attributes.

Dataset Used

The methodology was evaluated on the Pima Indians Diabetes Dataset (PIDD), which was taken from the UCI Repository. This dataset comprises of medical detail of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where value of one class ’0’ treated as tested negative for diabetes and value of another class ’1’ is treated as tested positive for diabetes.

Accuracy Measures

The algorithms were evaluated using internal cross-validation 10-folds. Accuracy, F-Measure, Recall, Precision and ROC (Receiver Operating Curve) measures were used for the classification of this work.

PERFORMANCE METRICS – MEASURE:

The performance of the machine learning classification algorithms used in the study was evaluated using several measures:

1. **Accuracy**: This measure determines the accuracy of the algorithm in predicting instances. It is calculated as the sum of True Positives (TP) and True Negatives (TN) divided by the total number of samples.

$$A = (TP + TN) / (\text{Total number of samples})$$

2. **Precision**: This measure evaluates the classifier's correctness or accuracy. It is calculated as the True Positives (TP) divided by the sum of True Positives and False Positives (FP).

$$P = TP / (TP + FP)$$

3. **Recall**: This measure is used to evaluate the classifier's completeness or sensitivity. It is calculated as the True Positives (TP) divided by the sum of True Positives and False Negatives (FN).

$$R = TP / (TP + FN)$$

4. **F-Measure**: This measure is the weighted average of precision and recall.

$$F = 2 * (P * R) / (P + R)$$

5. **ROC (Receiver Operating Curve)**: ROC curves are used to compare the usefulness of tests.

These measures were used to evaluate the performance of the Naive Bayes, SVM, and Decision Tree algorithms. The results showed that the Naive Bayes algorithm outperformed the others with the highest accuracy of 76.30%.

ALGORITHMS:

The study used three machine learning classification algorithms to predict the likelihood of diabetes in patients. These algorithms were:

1. **Support Vector Machine (SVM)**: SVM is a supervised machine learning model used in classification. It aims to find the best highest-margin separating hyperplane between two classes. The hyperplane should be selected which is far from the data points from each category. The points that lie nearest to the margin of the classifier are the support vectors.
2. **Naive Bayes Classifier**: Naive Bayes is a classification technique which assumes all features are independent and unrelated to each other. It is based on conditional probability and is considered a powerful algorithm for classification. It works well for data with imbalancing problems and missing values.
3. **Decision Tree Classifier**: Decision Tree is a supervised machine learning algorithm used to solve classification problems. It uses nodes and internodes for prediction and classification. Root nodes classify the instances with different features. In every stage, the decision tree chooses each node by evaluating the highest information gain among all the attributes.

The performance of these algorithms was evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. The Naive Bayes algorithm outperformed the others with the highest accuracy of 76.30%.

TEST BED:

The test bed for this study is the Pima Indians Diabetes Database (PIDD), sourced from the UCI machine learning repository. This dataset comprises of medical details of 768 instances which are female patients. The dataset also comprises numeric-valued 8 attributes where the value of one class '0' is treated as tested negative for diabetes and the value of another class '1' is treated as tested positive for diabetes.

Here is a brief description of the attributes in the dataset:

1. Number of times pregnant
2. Plasma glucose concentration

3. Diastolic blood pressure (mm Hg)
4. Skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. BMI (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age in years
9. Class ’0’ or ’1’

The experiments were performed using the WEKA tool, a software designed in New Zealand by the University of Waikato, which includes a collection of various machine learning methods for data classification, clustering, regression, visualization etc.

EXPERIMENT:

The experiment in the given text is about designing a model to predict the likelihood of diabetes in patients with maximum accuracy. The researchers used three machine learning classification algorithms: Decision Tree, SVM (Support Vector Machine), and Naive Bayes.

Dataset

The experiments were performed on the Pima Indians Diabetes Database (PIDD), sourced from the UCI machine learning repository. This dataset comprises medical details of 768 instances which are female patients. The dataset also includes numeric-valued 8 attributes where the value of one class '0' is treated as tested negative for diabetes and the value of another class '1' is treated as tested positive for diabetes.

Methodology

The performances of all the three algorithms were evaluated on various measures like Precision, Accuracy, F-Measure, and Recall. Accuracy was measured over correctly and incorrectly classified instances.

Results

The results obtained showed that the Naive Bayes algorithm outperformed the other algorithms with the highest accuracy of 76.30%. These results were verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

RESULTS:

The study used three machine learning classification algorithms, namely Decision Tree, SVM, and Naive Bayes, to predict the likelihood of diabetes in patients. The experiments were performed on the Pima Indians Diabetes Database (PIDD), sourced from the UCI machine learning repository.

The performance of the three algorithms was evaluated based on various measures such as Precision, Accuracy, F-Measure, and Recall. Accuracy was measured over correctly and incorrectly classified instances.

The results showed that the Naive Bayes algorithm outperformed the other two with the highest accuracy of 76.30%. These results were verified using Receiver Operating Characteristic (ROC) curves in a systematic manner.

Here is a summary of the results:

- **Naive Bayes**:
    - Precision: 0.759
    - Recall: 0.763
    - F-Measure: 0.760
    - Accuracy: 76.30%
    - ROC: 0.819
- **SVM**:
    - Precision: 0.424
    - Recall: 0.651
    - F-Measure: 0.513
    - Accuracy: 65.10%
    - ROC: 0.500
- **Decision Tree**:
    - Precision: 0.735
    - Recall: 0.738
    - F-Measure: 0.736
    - Accuracy: 73.82%
    - ROC: 0.751

INTERPRETATION OF DATA:

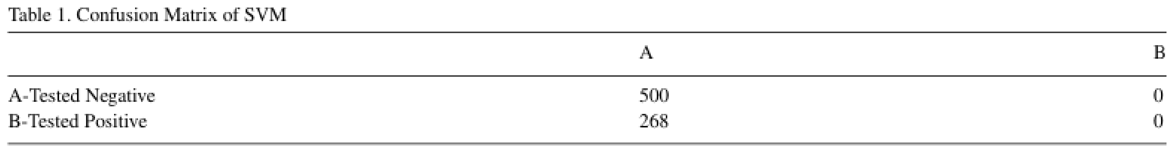


Table 1. Confusion Matrix of SVM

	A	B
A-Tested Negative	500	0
B-Tested Positive	268	0

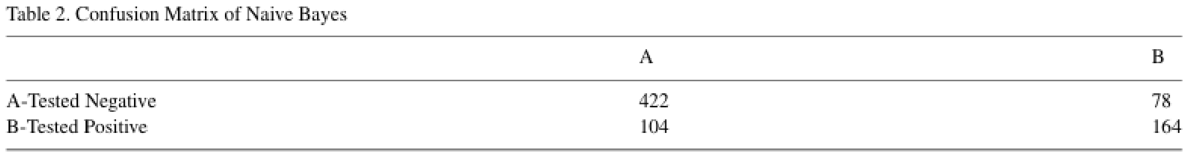


Table 2. Confusion Matrix of Naive Bayes

	A	B
A-Tested Negative	422	78
B-Tested Positive	104	164

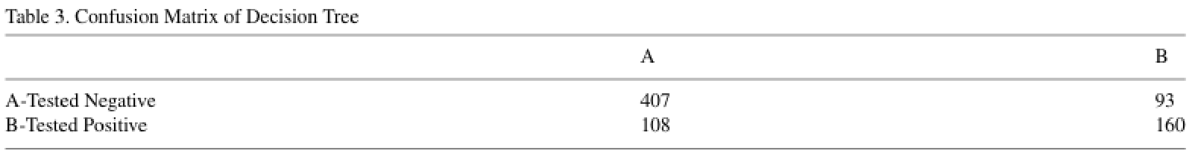


Table 3. Confusion Matrix of Decision Tree

	A	B
A-Tested Negative	407	93
B-Tested Positive	108	160

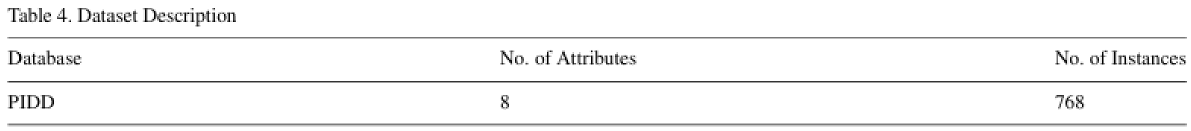


Table 4. Dataset Description

Database	No. of Attributes	No. of Instances
PIDD	8	768

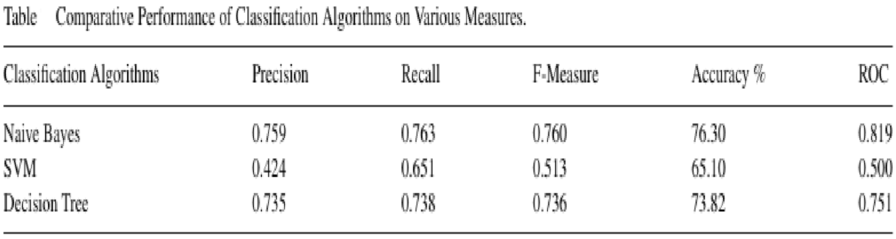


Table Comparative Performance of Classification Algorithms on Various Measures.

Classification Algorithms	Precision	Recall	F-Measure	Accuracy %	ROC
Naive Bayes	0.759	0.763	0.760	76.30	0.819
SVM	0.424	0.651	0.513	65.10	0.500
Decision Tree	0.735	0.738	0.736	73.82	0.751

Table represents different performance values of all classification algorithms calculated on various measures. From the Table, it is analyzed that Naïve Bayes showing the maximum accuracy. So the Naïve Bayes machine learning classifier can predict the chances of diabetes with more accuracy as compared to other classifiers.

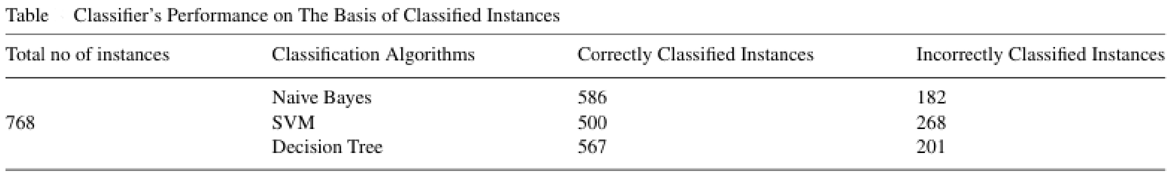


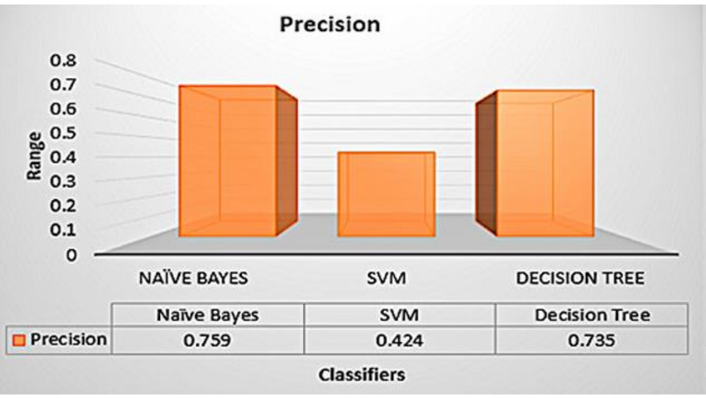
Table Classifier's Performance on The Basis of Classified Instances

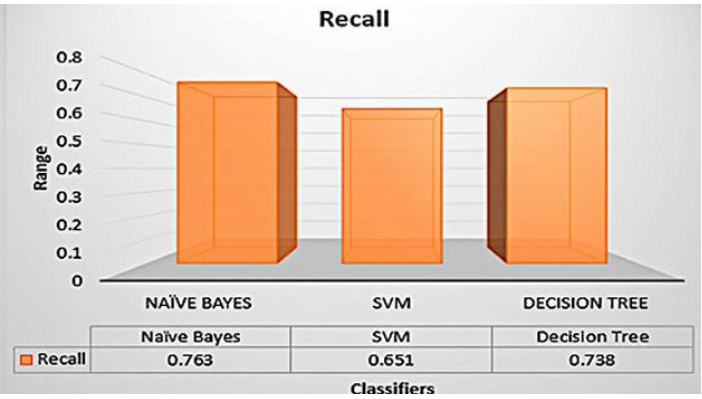
Total no of instances	Classification Algorithms	Correctly Classified Instances	Incorrectly Classified Instances
768	Naive Bayes SVM Decision Tree	586 500 567	182 268 201

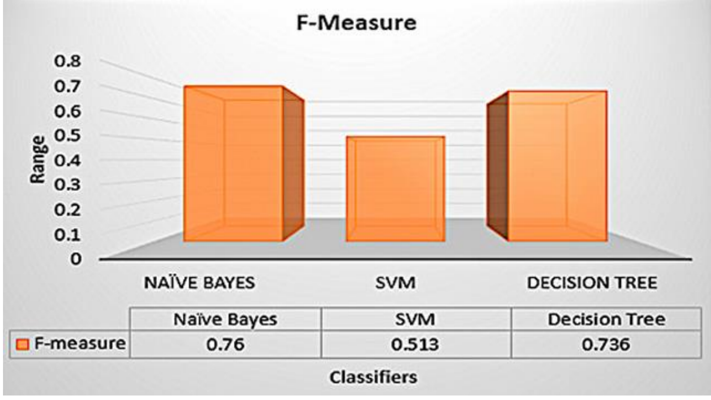
This table determines classifiers performance on the basis of classified instances. According to these classified instances, accuracy is calculated and analyzed. Performance of individual algorithm is evaluated on the basis of Correctly Classified Instances and Incorrectly Classified Instances out of a total number of instances.

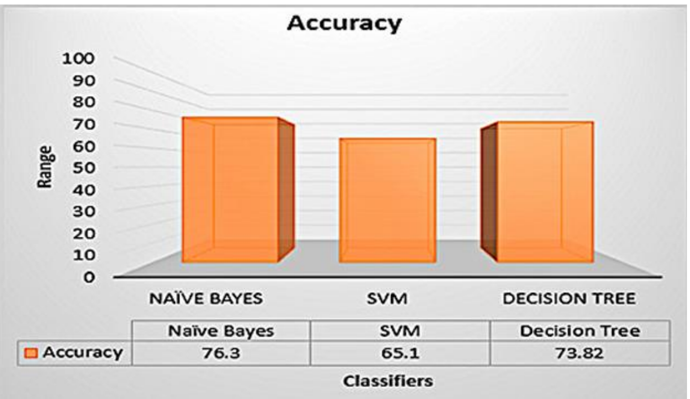
From past 2 tables, we can conclude that Naive Bayes classification algorithm outperforms comparatively other algorithms. So, Naive Bayes algorithm is considered as the best supervised machine learning method of this experiment because it gives higher accuracy in respective to other classification algorithms with an accuracy of 76.30 %.

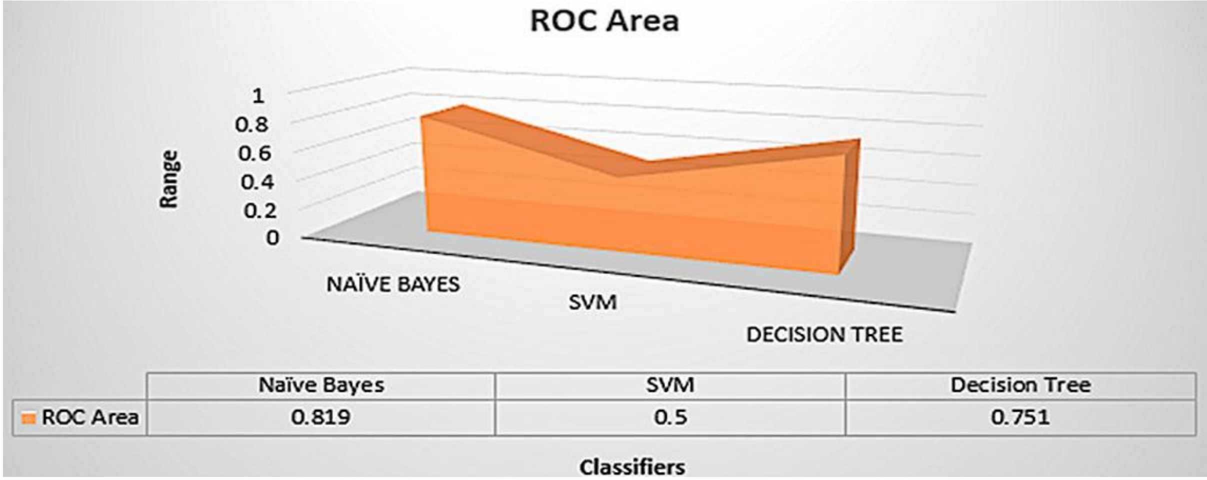
INTERPRETATION OF GRAPHS:

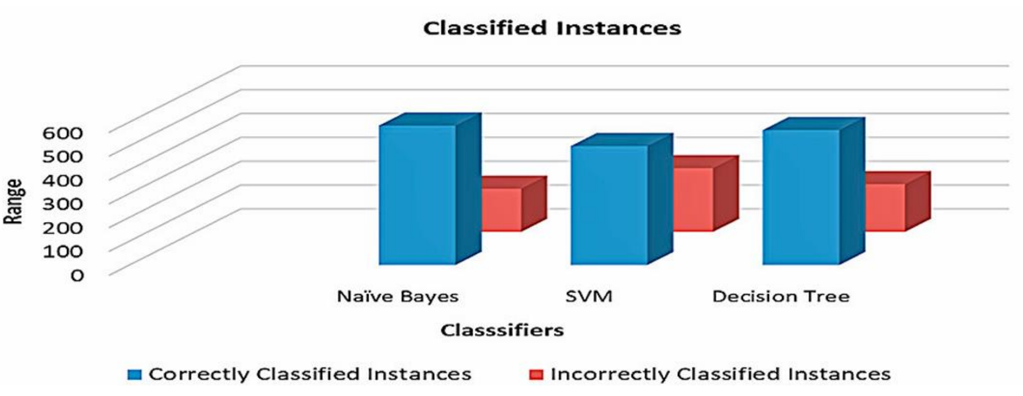












PROS:

1. **Early Detection**: Machine learning algorithms can help in the early detection of diabetes, which is crucial for managing the disease and preventing complications.
2. **High Accuracy**: The study found that the Naive Bayes classification algorithm had the highest accuracy (76.30%) among the tested algorithms.
3. **Use of Existing Data**: Machine learning algorithms can make use of existing medical data (like the Pima Indians Diabetes Database used in the study) to make predictions.
4. **Automation**: Once trained, these algorithms can make predictions automatically, reducing the need for manual analysis.

CONS:

1. **Data Quality**: The accuracy of predictions depends on the quality of the data used for training the algorithms. If the data is biased or incomplete, the predictions may also be inaccurate.
2. **Complexity**: Machine learning algorithms can be complex and require expertise to implement and interpret correctly.
3. **Lack of Explainability**: Some machine learning algorithms, like SVM, can be seen as a "black box" that makes predictions without easily understandable reasoning.
4. **Need for Continuous Training**: Machine learning models may need to be retrained as new data becomes available to maintain their accuracy.

FUTURE WORKS:

The future work of this study could involve the following:

- The designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. This would involve adapting the current model to different types of medical data and potentially incorporating additional machine learning algorithms to improve accuracy.
- The work can be extended and improved for the automation of diabetes analysis. This could involve developing a user-friendly interface for healthcare professionals to input patient data and receive a prediction of diabetes risk.
- The current model could be refined and improved. This could involve further tuning of the machine learning algorithms used, or the incorporation of additional patient data to improve the accuracy of predictions.

- The model could be tested on larger and more diverse datasets. This would provide a more robust validation of the model's performance and could potentially improve its predictive accuracy.
- The model could be integrated into a larger healthcare system. This could involve developing an API for the model that allows it to be easily integrated into existing healthcare IT systems. This would allow for real-time prediction of diabetes risk as part of routine patient care.

REFERENCES:

1. Aishwarya, R., Gayathri, P., Jaisankar, N., 2013. A Method for Classification Using Machine Learning Technique for Diabetes. International Journal of Engineering and Technology (IJET) 5, 2903–2908.
2. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.
3. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.
4. Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp. 181–184.
5. Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931 - 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427.
6. Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427.

PUBLISHED PAPER:

https://www.sciencedirect.com/science/article/pii/S1877050918308548?ref=pdf_download&fr=RR-2&rr=8688a683cb148a24

RESEARCH PAPER - 4:

Diabetes Prediction using Machine Learning Algorithms

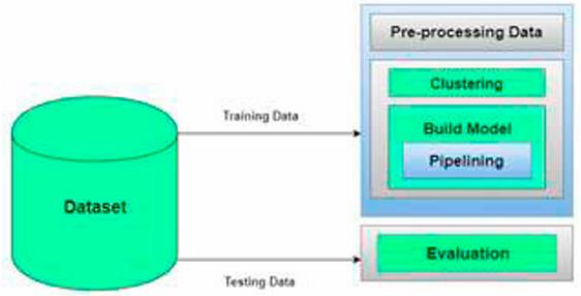
OBJECTIVE:

The objective of the paper is to propose a diabetes prediction model that improves the classification accuracy of diabetes diagnosis. The model incorporates several external factors responsible for diabetes, such as glucose levels, BMI, age, insulin levels, and lifestyle factors, along with regular factors. The authors aim to enhance the prediction accuracy by using machine learning algorithms and big data analytics, which can study large datasets and discover hidden patterns and information. The model is intended to improve the accuracy of classification, making it a valuable tool in healthcare industries.

PROPOSED METHODOLOGY:

The proposed methodology for diabetes prediction using machine learning algorithms consists of five different modules:

1. **Dataset Collection**: This module involves data collection and understanding the data to study the patterns and trends which helps in prediction and evaluating the results. The Diabetes dataset contains 800 records and 10 attributes.



2. **Data Pre-processing**: This phase handles inconsistent data to get more accurate and precise results. Missing values for selected attributes like Glucose level, Blood Pressure, Skin Thickness, BMI, and Age are imputed because these attributes cannot have zero values. Then the dataset is scaled to normalize all values.
3. **Clustering**: K-means clustering is implemented on the dataset to classify each patient into either a diabetic or non-diabetic class. Before performing K-means clustering, highly correlated attributes were found which were Glucose and Age. K-means clustering was performed on these two attributes.
4. **Model Building**: This is the most important phase which includes model building for prediction of diabetes. Various machine learning algorithms are implemented for diabetes prediction. These algorithms include Support Vector Classifier, Random Forest Classifier, Decision Tree Classifier, Extra

Tree Classifier, Ada Boost algorithm, Perceptron, Linear Discriminant Analysis algorithm, Logistic Regression, K-Nearest Neighbour, Gaussian Naïve Bayes, Bagging algorithm, Gradient Boost Classifier.

5. **Evaluation**: This is the final step of the prediction model. Here, the prediction results are evaluated using various evaluation metrics like classification accuracy, confusion matrix, and F1-score.

The proposed methodology aims to improve the accuracy and precision of diabetes prediction.

PERFORMANCE METRICS – MEASURE:

In the given text, the performance of various machine learning algorithms is evaluated using several metrics. These metrics provide a quantitative measure of the algorithms' effectiveness and accuracy. Here are the metrics mentioned:

1. **Classification Accuracy**: This is the ratio of the number of correct predictions to the total number of input samples.
2. **Confusion Matrix**: This is a table that describes the complete performance of the model. It includes four types of outcomes: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The accuracy for the matrix can be calculated by taking the average of the values lying across the main diagonal.
3. **F1 Score**: This is used to measure a test's accuracy. It is the harmonic mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).
4. **Precision**: This is the number of correct positive results divided by the number of positive results predicted by the classifier.
5. **Recall**: This is the number of correct positive results divided by the number of all relevant samples (all actual positives).

These metrics are used to evaluate the performance of the machine learning algorithms used in the study. The algorithm with the highest accuracy, precision, recall, and F1 score is considered the best performing algorithm.

ALGORITHMS:

The study used various machine learning algorithms to predict diabetes. These algorithms include:

1. **Support Vector Classifier (SVC)**

2. **Random Forest Classifier**
3. **Decision Tree Classifier**
4. **Extra Tree Classifier**
5. **AdaBoost Algorithm**
6. **Perceptron**
7. **Linear Discriminant Analysis (LDA)**
8. **Logistic Regression**
9. **K-Nearest Neighbour (KNN)**
10. **Gaussian Naive Bayes**
11. **Bagging Algorithm**
12. **Gradient Boost Classifier**

EXPERIMENT:

The experiment conducted in the text involves the application of various machine learning algorithms on a dataset to predict diabetes. The dataset contains 800 records and 10 attributes, including the number of pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, age, job type, and outcome.

The experiment is divided into five modules: Dataset Collection, Data Pre-processing, Clustering, Model Building, and Evaluation.

The experiment concludes with the finding that the Logistic Regression algorithm gives the highest accuracy of 96%. The application of a pipeline further improves the accuracy to 98.8% for the AdaBoost classifier.

RESULTS:

After applying various Machine Learning Algorithms on the dataset, the following accuracies were achieved:

- Decision Tree: 86%
- Gaussian NB: 93%
- LDA: 94%
- SVC: 60%
- Random Forest: 91%
- Extra Trees: 91%

- AdaBoost: 93%
- Perceptron: 76%
- Logistic Regression: 96%
- Gradient Boost Classifier: 93%
- Bagging: 90%
- KNN: 90%

The highest accuracy was achieved by the Logistic Regression algorithm, with an accuracy of 96%.

When using Pipelining, the highest accuracy achieved was 97.2% for Logistic Regression. The Pipelining results were as follows:

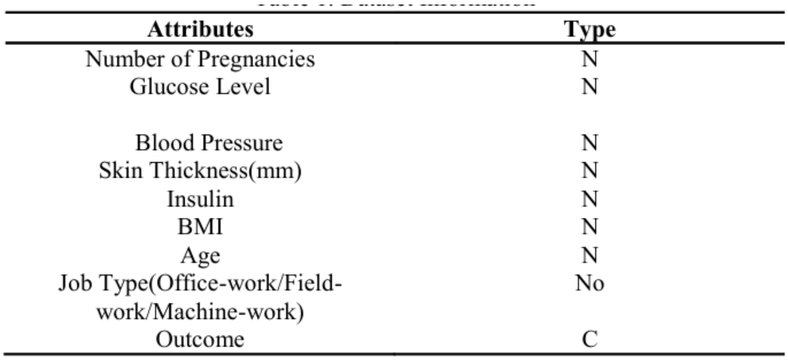
- AdaBoost Classifier: 98.8%
- Gradient Boost Classifier: 98.1%
- Random Forest Classifier: 98.1%
- Logistic Regression: 97.5%
- Extra Trees Classifier: 96.3%
- Linear Discriminant Analysis: 95%

The model improves accuracy and precision of diabetes prediction with this dataset compared to the existing dataset. The model with the highest accuracy was the AdaBoost classifier with an accuracy of 98.8%.

INTERPRETATION OF DATA:

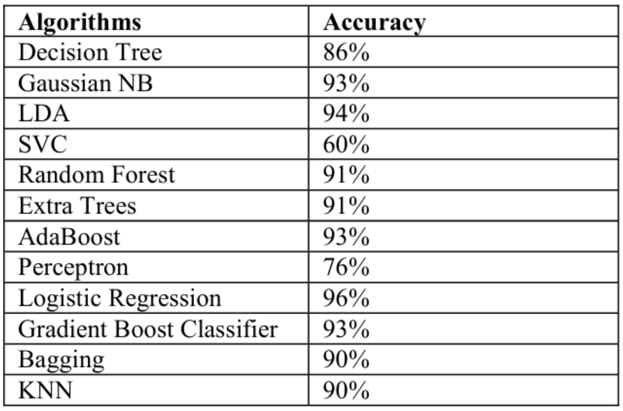
The research paper contains several tables that provide information about the dataset used, the accuracy of various machine learning algorithms, and the performance of these algorithms in predicting diabetes.

Table : Dataset Information - This table provides details about the dataset used in the study, which includes various attributes related to diabetes such as the number of pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, age, and job type. The 'Outcome' attribute indicates whether the individual has diabetes or not.



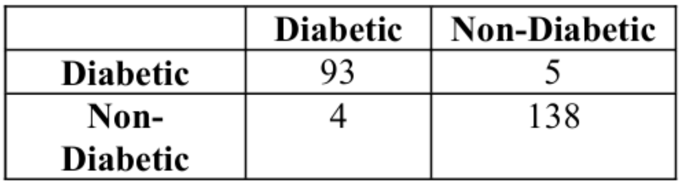
Attributes	Type
Number of Pregnancies	N
Glucose Level	N
Blood Pressure	N
Skin Thickness(mm)	N
Insulin	N
BMI	N
Age	N
Job Type(Office-work/Field-work/Machine-work)	No
Outcome	C

Table : Accuracy Table - This table presents the accuracy of various machine learning algorithms applied to the dataset. For instance, the Decision Tree algorithm has an accuracy of 86%, meaning it correctly predicted the outcome 86% of the time.



Algorithms	Accuracy
Decision Tree	86%
Gaussian NB	93%
LDA	94%
SVC	60%
Random Forest	91%
Extra Trees	91%
AdaBoost	93%
Perceptron	76%
Logistic Regression	96%
Gradient Boost Classifier	93%
Bagging	90%
KNN	90%

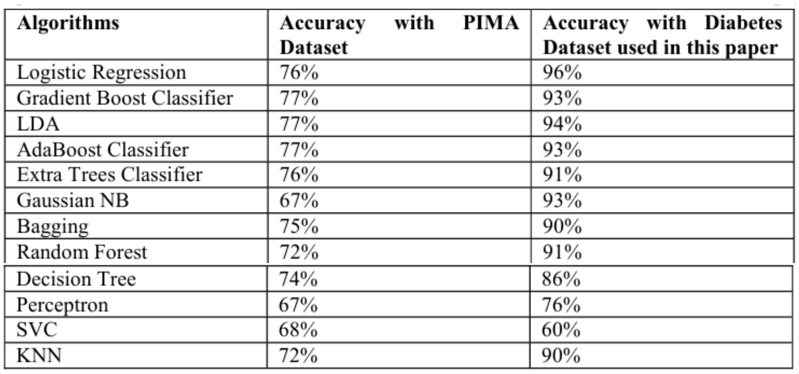
Table : Confusion Matrix for Logistic Regression - This table provides a detailed breakdown of the Logistic Regression algorithm's performance. It shows the number of true positives, false positives, false negatives, and true negatives.



	Diabetic	Non-Diabetic
Diabetic	93	5
Non-Diabetic	4	138

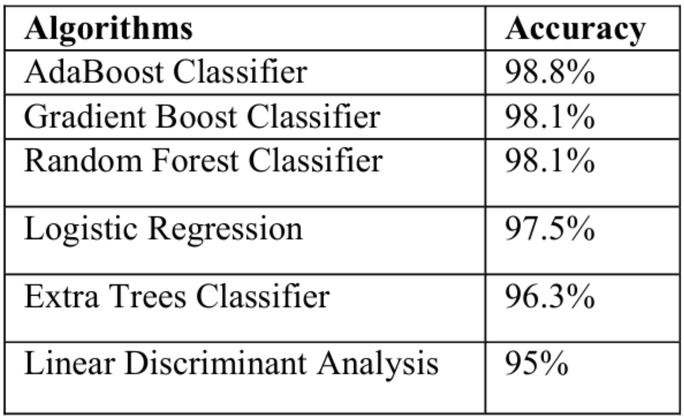
Table : Comparison between accuracies of PIMA Diabetes Dataset and Diabetes Dataset used in this study

This table compares the accuracy of various machine learning algorithms when applied to two different datasets: the PIMA Diabetes Dataset and the Diabetes Dataset used in this study.



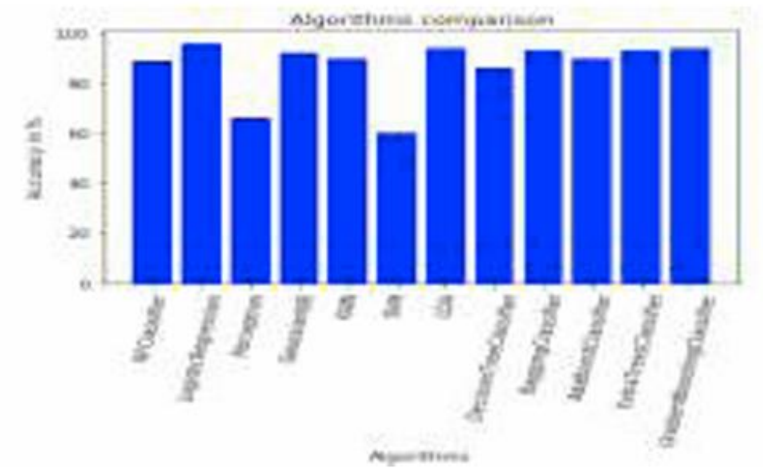
Algorithms	Accuracy with PIMA Dataset	Accuracy with Diabetes Dataset used in this paper
Logistic Regression	76%	96%
Gradient Boost Classifier	77%	93%
LDA	77%	94%
AdaBoost Classifier	77%	93%
Extra Trees Classifier	76%	91%
Gaussian NB	67%	93%
Bagging	75%	90%
Random Forest	72%	91%
Decision Tree	74%	86%
Perceptron	67%	76%
SVC	68%	60%
KNN	72%	90%

Table : Pipelining Results - This table shows the accuracy of various machine learning algorithms when used in a pipeline. The AdaBoost Classifier had the highest accuracy of 98.8% when used in a pipeline.



Algorithms	Accuracy
AdaBoost Classifier	98.8%
Gradient Boost Classifier	98.1%
Random Forest Classifier	98.1%
Logistic Regression	97.5%
Extra Trees Classifier	96.3%
Linear Discriminant Analysis	95%

INTERPRETATION OF GRAPHS:



The graph shows the comparison of various machine learning algorithms based on accuracies.

Classification has been done using various algorithms of which Logistic regression gives highest accuracy of 96. Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8.

PROS:

1. **Improved Accuracy**: The use of machine learning algorithms, particularly Logistic Regression, has shown to improve the accuracy of diabetes prediction up to 96%.
2. **Efficiency**: The proposed model uses big data analytics to study large datasets and find hidden patterns, which can lead to more efficient diagnosis and treatment.
3. **Comprehensive Analysis**: The model takes into account a variety of factors that can cause Diabetes Mellitus.
4. **Use of Multiple Algorithms**: The model uses various machine learning algorithms, which can provide a more comprehensive and reliable prediction.

CONS:

1. **Data Quality**: The accuracy of the model heavily depends on the quality and completeness of the data.
2. **Time-Consuming Preprocessing**: The data preprocessing phase can be time-consuming.
3. **Complexity**: The model uses a variety of machine learning algorithms and techniques, which can be complex to implement.
4. **Need for Regular Updates**: The model needs to be regularly updated with new data to maintain its accuracy and reliability.

FUTURE WORKS:

The future work for this study could involve extending the model to predict how likely non-diabetic people can develop diabetes in the next few years, incorporating more external factors that could influence the onset of diabetes, testing the model with more diverse datasets, exploring other machine learning algorithms or techniques, and developing a user-friendly interface or application for healthcare professionals.

REFERENCES:

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar," Predictive Analysis of Diabetic Patient Data Using machine Learning and Hadoop", International Conference On I-SMAC, 978-1-5090-3243-32017.

[2] Ayush Anand and Divya Shakti," Prediction of Diabetes Based on Personal Lifestyle Indicators", 1st International Conference on Next Generation Computing Technologies, 978-1-4673-6809-4, September 2015.

[3] B. Nithya and Dr. V. Ilango," Predictive Analytics in Health Care Using Machine Learning Tools and Techniques", International Conference on Intelligent Computing and Control Systems, 978-1-5386-2745-7,2017.

[4] Dr Saravana kumar N M, Eswari T, Sampath P and Lavanya S," Predictive Methodology for Diabetic Data Analysis in Big Data", 2nd International Symposium on Big Data and Cloud Computing, 2015.

[5] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly," Diagnosis of Diabetes Using Classification Mining Techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[6] P. Suresh Kumar and S. Pranavi "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics", International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20, 2017.

[7] Mani Butwall and Shraddha Kumar," A Data Mining Approach for the Diagnosis of Diabetes Mellitus using Random Forest Classifier", International Journal of Computer Applications, Volume 120 - Number 8,2015.

[8] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis", International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.

PUBLISHED PAPER:

https://www.sciencedirect.com/science/article/pii/S1877050920300557

RESEARCH PAPER - 5:

Improving the Accuracy of Diabetes Diagnosis Applications through a Hybrid Feature Selection Algorithm

OBJECTIVE:

The objective of the research paper is to improve the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm. The paper focuses on diagnosing a diabetic patient through data mining techniques. The proposed method has three steps: preprocessing, feature selection, and classification. The researchers aim to diagnose diabetes in the early stages of the disease, which can help patients to stay home and care for their health, reducing the risk of being infected with the coronavirus. The study also aims to increase the reliability of the cure and decision making in healthcare by developing useful systems and algorithms.

PROPOSED METHODOLOGY:

The proposed methodology in the given text involves the use of artificial intelligence and data mining techniques to improve the accuracy of diabetes diagnosis applications.

1. **Preprocessing**: This step involves preparing the data for analysis. The text does not provide specific details on what this preprocessing entails, but it typically involves cleaning the data, handling missing values, and normalizing the data.

2. **Feature Selection**: This step involves identifying the most relevant features (or variables) in the data that will be used for the diagnosis. The authors propose a hybrid feature selection algorithm that combines several combinations of the Harmony search algorithm, genetic algorithm, and particle swarm optimization algorithm with K-means clustering. These combinations have not been examined before for diabetes diagnosis applications.

3. **Classification**: After the features have been selected, the K-nearest neighbor algorithm is used for classification of the diabetes dataset. This involves assigning each data point to a particular class or category based on its features.

The authors evaluate the performance of their proposed method using measures such as sensitivity, specificity, and accuracy. The results indicate that their method, with an accuracy of 91.65%, outperforms the results of earlier methods examined in the article.

PERFORMANCE METRICS – MEASURE:

The performance of the proposed hybrid feature selection algorithms in the study was evaluated using three widely used parameters: accuracy, sensitivity, and specificity.

1. **Accuracy**: This is the proportion of true results (both true positives and true negatives) in the total dataset. It is calculated using the formula:
2. Accuracy = (TP + TN) / (TP + TN + FP + FN)

where TP is the rate of records that have diabetes and are correctly classified as diabetic people, TN is the rate of records which do not have diabetes and are correctly classified as non-diabetic people, FP is the rate of records that are diabetics but are incorrectly classified as non-diabetics, and FN is the rate of non-diabetics that are incorrectly classified as diabetics.

3. **Sensitivity (Recall)**: This is the ability of a test to correctly identify positive results to get the proportion of actual positives which are correctly identified. It is calculated using the formula:
4. Recall = TP / (TP + FP)
5. **Specificity (Precision)**: This is the ability of the test to correctly identify negative results to get the proportion of actual negatives which are correctly identified. It is calculated using the formula:
6. Precision = TP / (TP + FP)

The tenfold cross-validation method was used to train and test the models. The results of the different examinations were then compared to evaluate the performance of the proposed method.

ALGORITHMS:

The text discusses several algorithms used in the study for diagnosing diabetes. These algorithms are primarily used for feature selection and classification.

Metaheuristic and Data Mining Algorithms

1. **Genetic Algorithm (GA)**: This is a metaheuristic search optimization algorithm based on Darwin’s survival theory. It starts by generating an initial random population and then generates a series of new populations. The algorithm ends when one of the terminating conditions, time limit or fitness limits, are satisfied.
2. **Harmony Search Algorithm**: This optimization algorithm is inspired by the work of musicians to enhance the instrument’s performance. It works similarly to GA in generating the next Harmony generation based on the

current population. The goal of Harmony search algorithm is finding the best response from a set of responses.

3. **Particle Swarm Optimization (PSO)**: This algorithm is inspired by fish shoaling and bird flocking social behavior. It works by determining the current position of a particle, the pbest of the particle, the gbest of the group, and the velocity of the particle.
4. **K-Nearest Neighbors (KNN)**: This algorithm is used for classification of the diabetes dataset. It works by determining the number K of the neighbors, computing the distances between the desired data-point and its K neighbors using the Euclidean distance, selecting the K nearest neighbors in terms of Euclidean distance calculated, and counting the number of data-points of each class from the k neighbors selected.
5. **K-means Clustering Algorithm**: This algorithm is used to divide the whole dataset into two clusters. It works by setting k random points as means, placing each item into a group of items with the closest average and updating it to include the new item, and repeating the process to meet the stopping criteria.

Proposed Hybrid Algorithms

1. **GA-Kmeans**: This is a combination of GA and K-means clustering. First, clustering is performed, and all of the dataset records are separated into two clusters. Then, GA is used to investigate the relationship between the records in each cluster and to determine the impact of each feature to assign a record to a cluster.
2. **GA-PSO-Kmeans**: This is a combination of GA, PSO, and K-means for feature selection. GA and PSO are used together to decrease the chance of getting stuck in local minima.
3. **HR-Kmeans**: This is a combination of Harmony search algorithm and K-means clustering algorithm for feature selection. The Harmony search algorithm is used to find proper features from each of the clusters provided by the K-means algorithm.

TEST BED:

The test bed in this study is the PIMA Indian diabetes dataset. This dataset is used to assess the proposed method for diagnosing Type 2 diabetes. It is a non-linear dataset prepared from Indian women aged 21 years or older, and is available in the UCI's machine learning repository. It includes 768 records; each record is defined with eight integer-real attributes. There is another attribute, which is the label with values of 1 indicating patients that have diabetes and 0 for those that do not.

EXPERIMENT:

The experiment conducted in the study involved using the PIMA Indian diabetes dataset to assess the proposed method for diagnosing Type 2 diabetes. The dataset is non-linear and includes 768 records, each defined with eight integer-real attributes and a label indicating whether the individual has diabetes (1) or not (0).

Three hybrid feature selection algorithms were proposed and applied to the dataset:

1. **GA-Kmeans:** This combination involved using the Genetic Algorithm (GA) and K-means clustering. The dataset records were separated into two clusters using K-means, then GA was used to investigate the relationship between the records in each cluster and determine the impact of each feature to assign a record to a cluster.

2. **GA-PSO-Kmeans:** This combination involved using GA, Particle Swarm Optimization (PSO), and K-means clustering. Similar to the GA-Kmeans combination, K-means was used for clustering, and then GA and PSO were used to find relationships between records and rank the features for classification.

3. **HR-Kmeans:** This combination involved using the Harmony search algorithm and K-means clustering. Again, K-means was used for clustering, and then the Harmony search algorithm was used to find proper features from each of the clusters.

After the feature selection stage, K-Nearest Neighbors (KNN) was employed for the classification of the diabetes records. The accuracy of the classification was then evaluated. The best accuracy of 91.65% was obtained using the HR-Kmeans hybrid.

RESULTS:

The results of the study are summarized as follows:

1. **Standard Algorithms**: The standard Support Vector Machine (SVM), Decision Tree (DT), and K-Nearest Neighbors (KNN) algorithms were first applied to the PIMA Indian diabetes dataset without feature selection. The accuracies achieved were 82.85%, 84.30%, and 86.63% respectively.

2. **Hybrid Algorithms**: Three hybrid feature selection algorithms were then applied to the dataset: GA-Kmeans, GA-PSO-Kmeans, and HR-Kmeans. The best features selected by HR-Kmeans included BloodPressure, Glucose, and Insulin. The accuracies achieved by these hybrid algorithms were 88.02%, 89.64%, and 91.65% respectively, with HR-Kmeans achieving the highest accuracy.

3. **Comparison with Previous Studies**: The results of the proposed hybrid algorithms were compared with the results reported in previous studies. The

proposed HR-Kmeans hybrid algorithm achieved a higher accuracy (91.65%) than the best previously reported results.

4. **Feature Selection**: The features selected by the best-performing hybrid algorithm (HR-Kmeans) were BloodPressure, Glucose, and Insulin. In addition to these, Age and BMI were also selected by the GA-Kmeans and GA-PSO-Kmeans combinations.

5. **Processing Time**: The model processing time for the best combination of algorithms (HR-Kmeans) was 24.44 seconds. For GA-Kmeans and GA-PSO-Kmeans, the model processing times were 22.3 seconds and 24.43 seconds respectively.

In conclusion, the study found that the proposed hybrid feature selection algorithms, particularly the HR-Kmeans algorithm, improved the accuracy of diabetes diagnosis using the PIMA Indian diabetes dataset.

INTERPRETATION OF DATA:

Table 1 provides a description of the PIMA Indian diabetes dataset, which includes eight attributes for each record: number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The ninth attribute is the class label, indicating whether the person has diabetes or not.

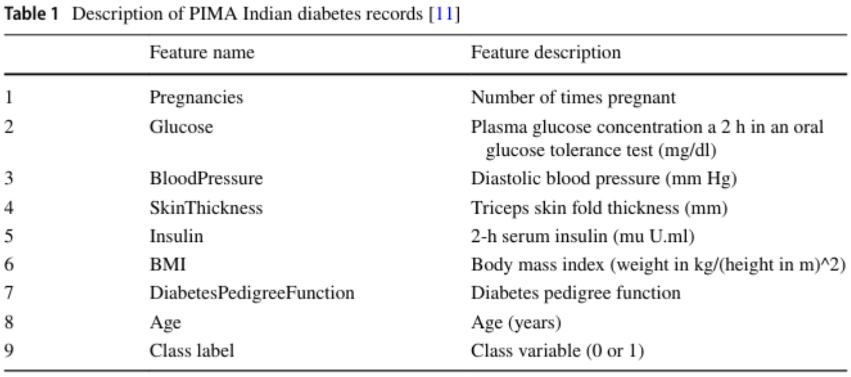


Table 1 Description of PIMA Indian diabetes records [11]

	Feature name	Feature description
1	Pregnancies	Number of times pregnant
2	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test (mg/dl)
3	BloodPressure	Diastolic blood pressure (mm Hg)
4	SkinThickness	Triceps skin fold thickness (mm)
5	Insulin	2-h serum insulin (mu U.ml)
6	BMI	Body mass index (weight in kg/(height in m)^2)
7	DiabetesPedigreeFunction	Diabetes pedigree function
8	Age	Age (years)
9	Class label	Class variable (0 or 1)

Table 2 :This table presents the results of different feature selection and classification methods applied to the PIMA Indian diabetes dataset. The methods include standard classifiers (SVM, DT, KNN) and hybrid methods (GA-Kmeans, GA-PSO-Kmeans, HR-Kmeans). The table shows the sensitivity, specificity, and accuracy of each method. The HR-Kmeans method achieved the highest accuracy of 91.65%.

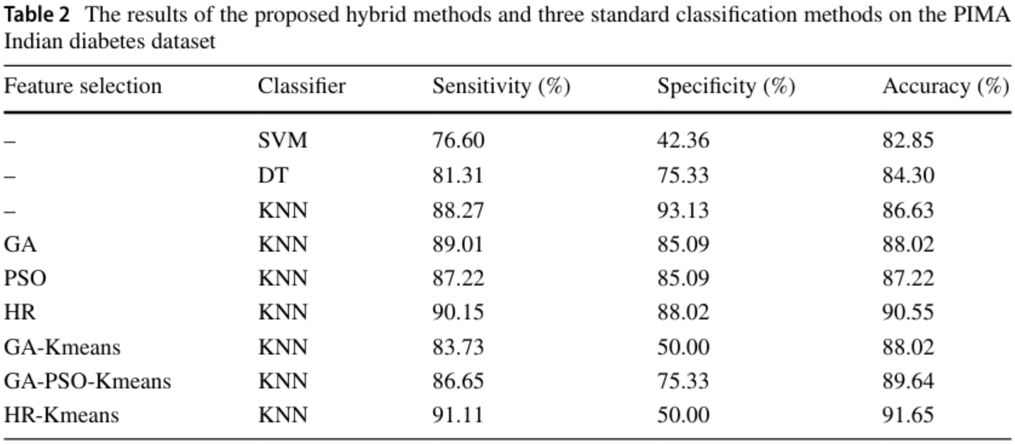
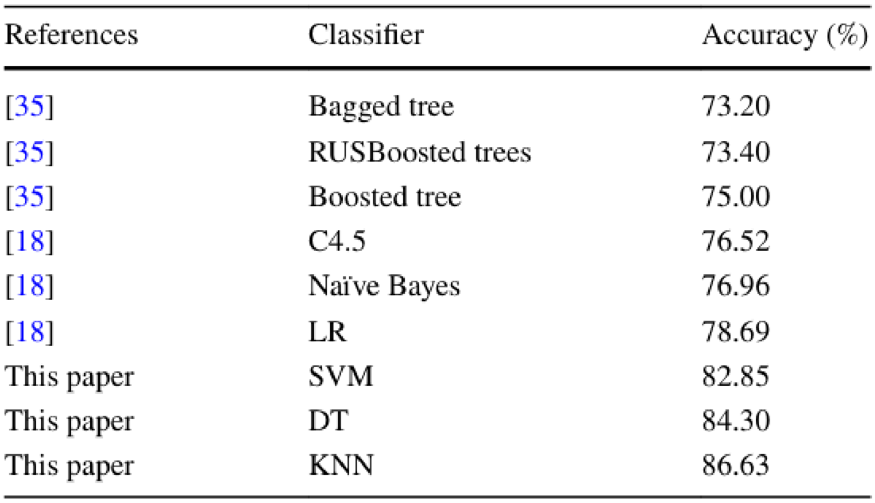


Table 2 The results of the proposed hybrid methods and three standard classification methods on the PIMA Indian diabetes dataset

Feature selection	Classifier	Sensitivity (%)	Specificity (%)	Accuracy (%)
–	SVM	76.60	42.36	82.85
–	DT	81.31	75.33	84.30
–	KNN	88.27	93.13	86.63
GA	KNN	89.01	85.09	88.02
PSO	KNN	87.22	85.09	87.22
HR	KNN	90.15	88.02	90.55
GA-Kmeans	KNN	83.73	50.00	88.02
GA-PSO-Kmeans	KNN	86.65	75.33	89.64
HR-Kmeans	KNN	91.11	50.00	91.65

Table 3 The comparisons of the accuracy of different standard classifiers examined in this paper and reported in previous studies



References	Classifier	Accuracy (%)
[35]	Bagged tree	73.20
[35]	RUSBoosted trees	73.40
[35]	Boosted tree	75.00
[18]	C4.5	76.52
[18]	Naïve Bayes	76.96
[18]	LR	78.69
This paper	SVM	82.85
This paper	DT	84.30
This paper	KNN	86.63

Table 3 compares the accuracy of different standard classifiers used in this study and those reported in previous studies. The KNN classifier used in this study achieved the highest accuracy of 86.63%.

Table 4 compares the accuracy of different hybrid algorithms proposed in previous researches with the hybrid algorithms proposed in this study. The HR-Kmeans hybrid algorithm proposed in this study achieved the highest accuracy of 91.65%.

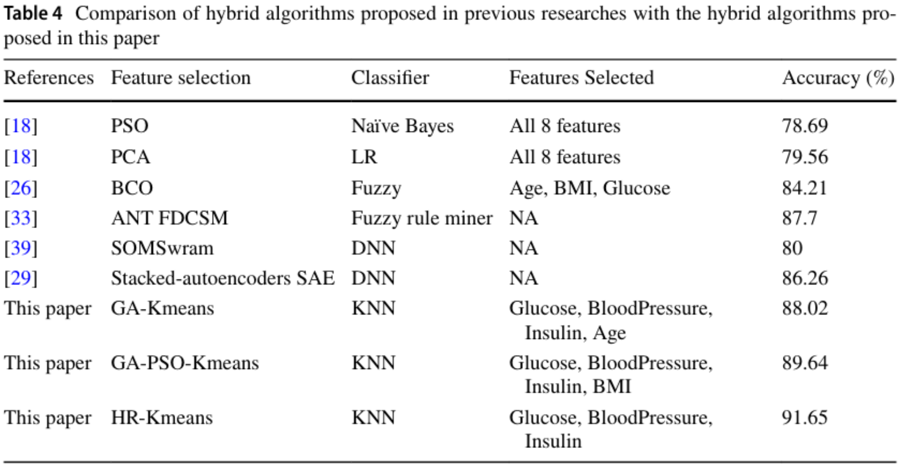
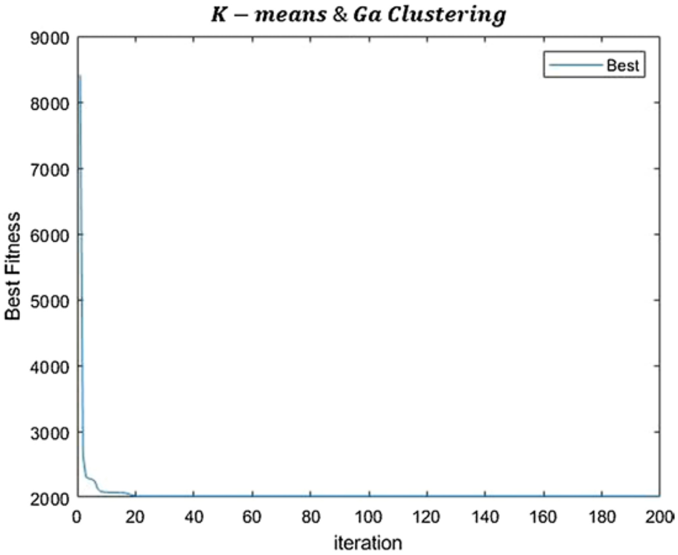


Table 4 Comparison of hybrid algorithms proposed in previous researches with the hybrid algorithms proposed in this paper

References	Feature selection	Classifier	Features Selected	Accuracy (%)
[18]	PSO	Naïve Bayes	All 8 features	78.69
[18]	PCA	LR	All 8 features	79.56
[26]	BCO	Fuzzy	Age, BMI, Glucose	84.21
[33]	ANT FDCSM	Fuzzy rule miner	NA	87.7
[39]	SOMSwram	DNN	NA	80
[29]	Stacked-autoencoders SAE	DNN	NA	86.26
This paper	GA-Kmeans	KNN	Glucose, BloodPressure, Insulin, Age	88.02
This paper	GA-PSO-Kmeans	KNN	Glucose, BloodPressure, Insulin, BMI	89.64
This paper	HR-Kmeans	KNN	Glucose, BloodPressure, Insulin	91.65

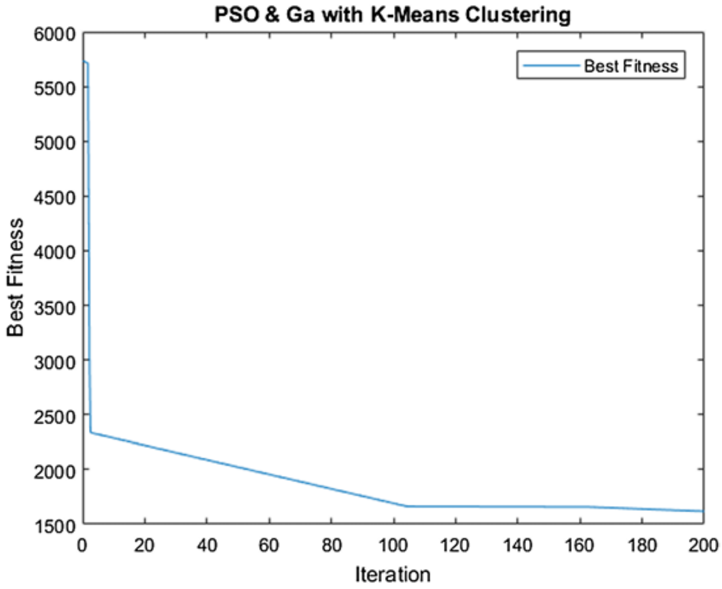
INTERPRETATION OF GRAPHS:



i)The graph provided represents the performance of a hybrid algorithm that combines Genetic Algorithms (GA) and K-means clustering over a series of iterations. The graph shows the Decreasing fitness function by the GA-Kmeans hybrid and is labeled "K-means & GA Clustering.”

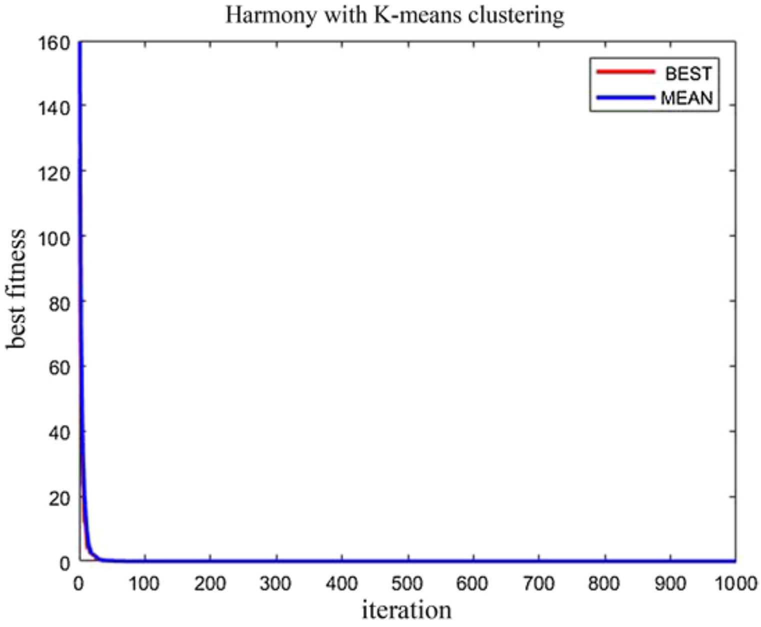
This pattern suggests that the GA-Kmeans hybrid algorithm is effective at quickly finding a good clustering solution and then refining it over time. The rapid initial decrease could be due to the GA's global search capabilities, which help to escape local optima, while the K-means algorithm fine-tunes the solution by iteratively improving the cluster assignments. This indicates that the algorithm is converging to a solution and making only minor improvements in the later iterations.

ii)The graph shows the Decreasing fitness function by GA-PSO-Kmeans hybrid, represents the performance of a hybrid algorithm that combines Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and K-Means clustering over a series of iterations.



- The best fitness value starts at a high level, indicating a lower quality solution, and rapidly decreases as the number of iterations increases. This suggests that the hybrid algorithm is effectively optimizing a problem, finding better solutions as it iterates.
- The most significant improvements appear to occur in the early iterations, with the curve flattening out as it progresses. This indicates that the algorithm is converging towards an optimal or near-optimal solution.
- By the 200th iteration, the best fitness value has settled to a level just above 1500. This implies that the algorithm's performance has stabilized and further iterations may not result in significant improvements.

iii) The graph shows the Decreasing fitness function by HR-Kmeans hybrid, representing the performance of a hybrid algorithm that combines Harmony Search (HS) with K-means clustering over a series of iterations.



The graph includes two lines, one representing the "BEST" fitness value and the other representing the "MEAN" fitness value across iterations. The

"BEST" line (in blue) shows a sharp decrease in the fitness value at the very beginning, suggesting that the algorithm quickly finds a good solution. After this initial drop, the "BEST" line flattens out, indicating that subsequent iterations do not significantly improve the best-found solution. The "MEAN" line (in red) also shows a decrease, but it is less steep than the "BEST" line. This suggests that on average, the fitness values across different runs or different solutions being considered by the algorithm are also improving, but not as dramatically as the best case. The "MEAN" line also flattens out, which implies that the average performance of the algorithm stabilizes as iterations increase.

Overall, the graph suggests that the HR-Kmeans hybrid algorithm is effective at quickly finding a good clustering solution and that the quality of this solution does not improve much after the initial iterations. This could indicate that the algorithm converges to a solution early on.

PROS:

1. The proposed method using a combination of metaheuristic algorithms and K-means clustering algorithm for feature selection improved the accuracy of diabetes diagnosis to 91.65%.
2. The use of artificial intelligence and data mining techniques can help in early diagnosis of diseases like diabetes, leading to better patient care and management.
3. The Harmony search algorithm used in the study achieved better accuracies than other metaheuristic algorithms and had better model processing time than those of Genetic Algorithm (GA) and Particle Swarm Optimization (PSO).
4. The study highlights the importance of such techniques in situations like the coronavirus pandemic, where healthcare workers are overworked due to a massive increase in the number of patients.

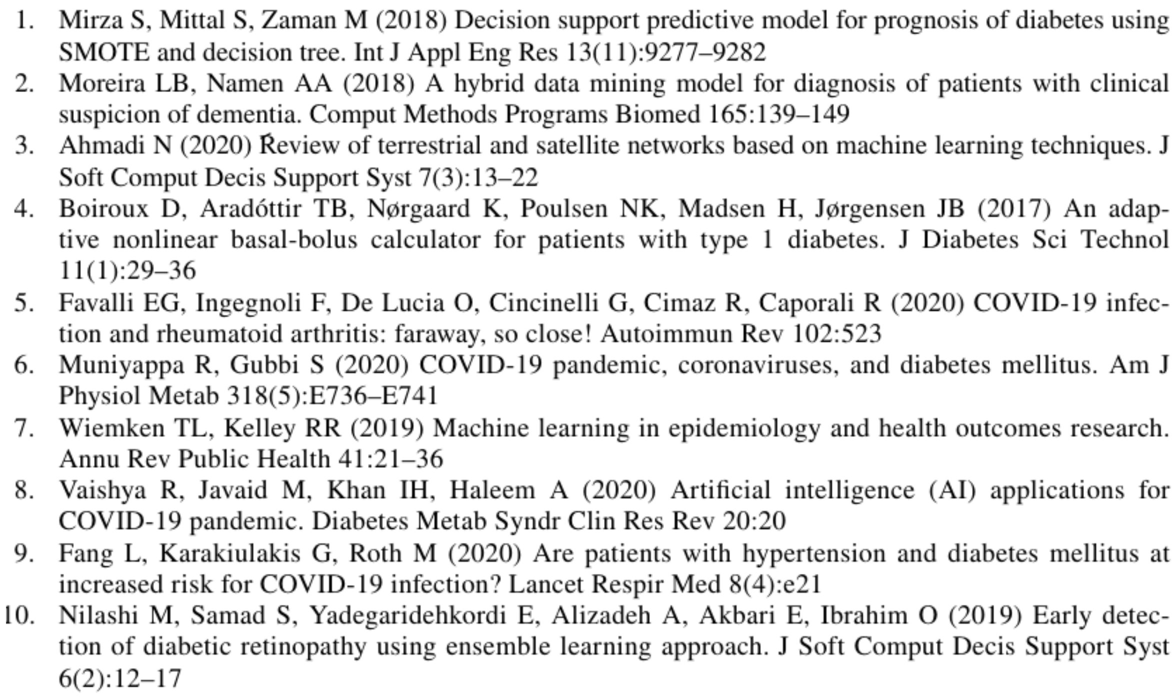
CONS:

1. The Harmony search algorithm does not work well with high-dimensional data. However, the dataset used in this study was not high-dimensional.
2. The results and effectiveness of the proposed method are highly dependent on the dataset used. The study used the PIMA Indian diabetes dataset, and the results might vary with different datasets.
3. The study lacks a valid comparison of the model processing time with other methods as most of the previous studies did not report their model processing time.
4. The use of multiple algorithms and techniques might increase the complexity of the system.

FUTURE WORKS:

1. Applying the proposed algorithms to local diabetes data.
2. Examining other combinations of metaheuristic algorithms.
3. Proposing new fitness functions for each of the heuristic algorithms for better feature rankings.
4. Evaluating the proposed method through mathematical and statistical tests such as McNamara's test.

REFERENCES:



1. Mirza S, Mittal S, Zaman M (2018) Decision support predictive model for prognosis of diabetes using SMOTE and decision tree. Int J Appl Eng Res 13(11):9277–9282
2. Moreira LB, Namen AA (2018) A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. Comput Methods Programs Biomed 165:139–149
3. Ahmadi N (2020) Ŕeview of terrestrial and satellite networks based on machine learning techniques. J Soft Comput Decis Support Syst 7(3):13–22
4. Boiroux D, Aradóttir TB, Nørgaard K, Poulsen NK, Madsen H, Jørgensen JB (2017) An adaptive nonlinear basal-bolus calculator for patients with type 1 diabetes. J Diabetes Sci Technol 11(1):29–36
5. Favalli EG, Ingegnoli F, De Lucia O, Cincinelli G, Cimaz R, Caporali R (2020) COVID-19 infection and rheumatoid arthritis: faraway, so close! Autoimmun Rev 102:523
6. Muniyappa R, Gubbi S (2020) COVID-19 pandemic, coronaviruses, and diabetes mellitus. Am J Physiol Metab 318(5):E736–E741
7. Wiemken TL, Kelley RR (2019) Machine learning in epidemiology and health outcomes research. Annu Rev Public Health 41:21–36
8. Vaishya R, Javaid M, Khan IH, Haleem A (2020) Artificial intelligence (AI) applications for COVID-19 pandemic. Diabetes Metab Syndr Clin Res Rev 20:20
9. Fang L, Karakiulakis G, Roth M (2020) Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? Lancet Respir Med 8(4):e21
10. Nilashi M, Samad S, Yadegaridehkordi E, Alizadeh A, Akbari E, Ibrahim O (2019) Early detection of diabetic retinopathy using ensemble learning approach. J Soft Comput Decis Support Syst 6(2):12–17

PUBLISHED PAPER:

https://link.springer.com/article/10.1007/s11063-021-10491-0

THE END