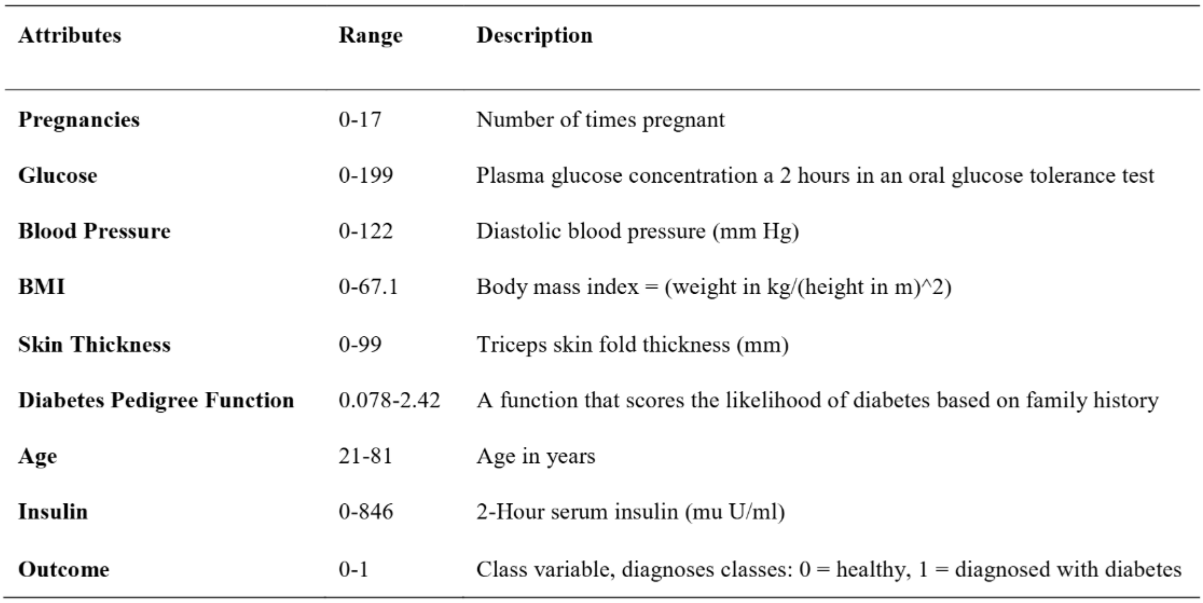
Analysis of Graphs and Tables in

"Diabetes Detection Using Machine Learning Classification Methods"

The paper presents several visual elements to communicate the dataset characteristics and model performance metrics.

Table 1: PIMA Indian Dataset Attributes Description



Attributes	Range	Description
Pregnancies	0-17	Number of times pregnant
Glucose	0-199	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Blood Pressure	0-122	Diastolic blood pressure (mm Hg)
BMI	0-67.1	Body mass index = (weight in kg/(height in m)^2)
Skin Thickness	0-99	Triceps skin fold thickness (mm)
Diabetes Pedigree Function	0.078-2.42	A function that scores the likelihood of diabetes based on family history
Age	21-81	Age in years
Insulin	0-846	2-Hour serum insulin (mu U/ml)
Outcome	0-1	Class variable, diagnoses classes: 0 = healthy, 1 = diagnosed with diabetes

This table provides essential information about the dataset's attributes, including:

- **Range**: The minimum and maximum values for each attribute
- **Description**: A brief explanation of what each attribute represents

The table effectively documents all nine attributes, including eight predictor variables and one target variable (Outcome). It shows the diversity of medical measurements used for prediction, from basic demographic information like age to specific clinical measurements like glucose levels and insulin.

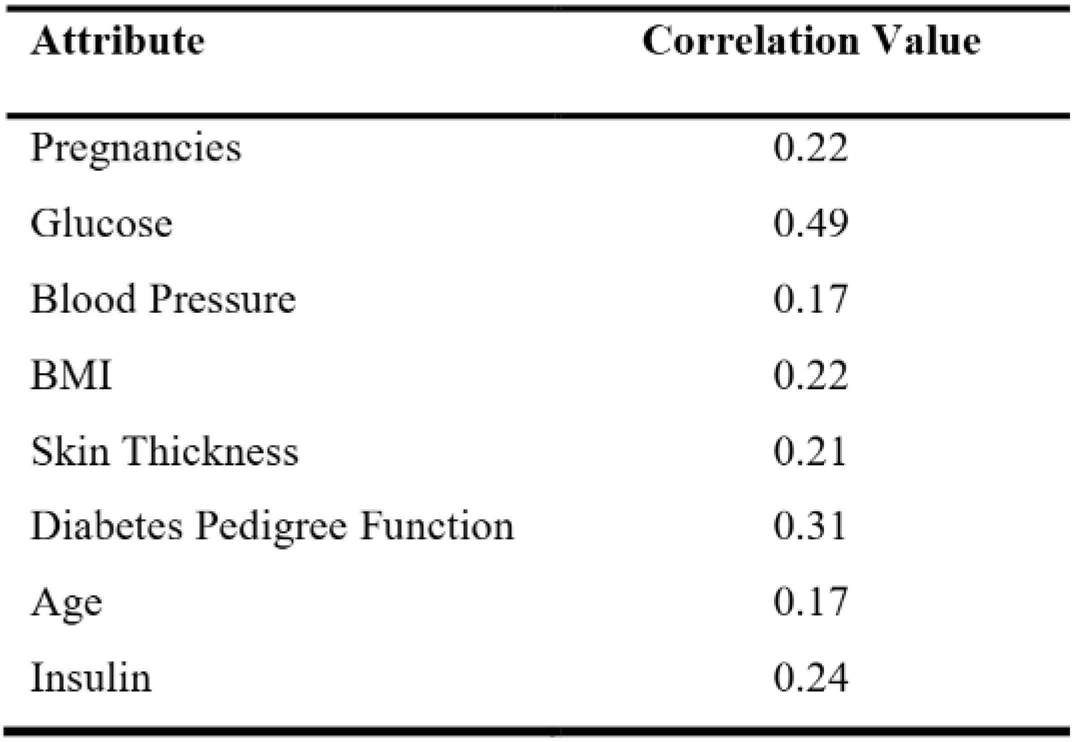
The table serves as a foundational reference point for understanding the dataset, showing reasonable ranges for each attribute (e.g., pregnancies ranging from 0-17, ages from 21-81). The clear descriptions help readers understand medical terminology like "Diabetes Pedigree Function."

Suggested improvements:

- Include the units of measurement for all attributes consistently (mm Hg is included for blood pressure but units aren't specified for other measurements)

- Add statistical information like mean and standard deviation to provide a better sense of the distribution
- Include information about missing values or zeros in the original dataset since this becomes important later

Table 2: Correlation with Outcome (Target)



Attribute	Correlation Value
Pregnancies	0.22
Glucose	0.49
Blood Pressure	0.17
BMI	0.22
Skin Thickness	0.21
Diabetes Pedigree Function	0.31
Age	0.17
Insulin	0.24

This table quantifies the relationship between each predictor variable and the target outcome (diabetes diagnosis) using Pearson correlation coefficients. The key insights include:

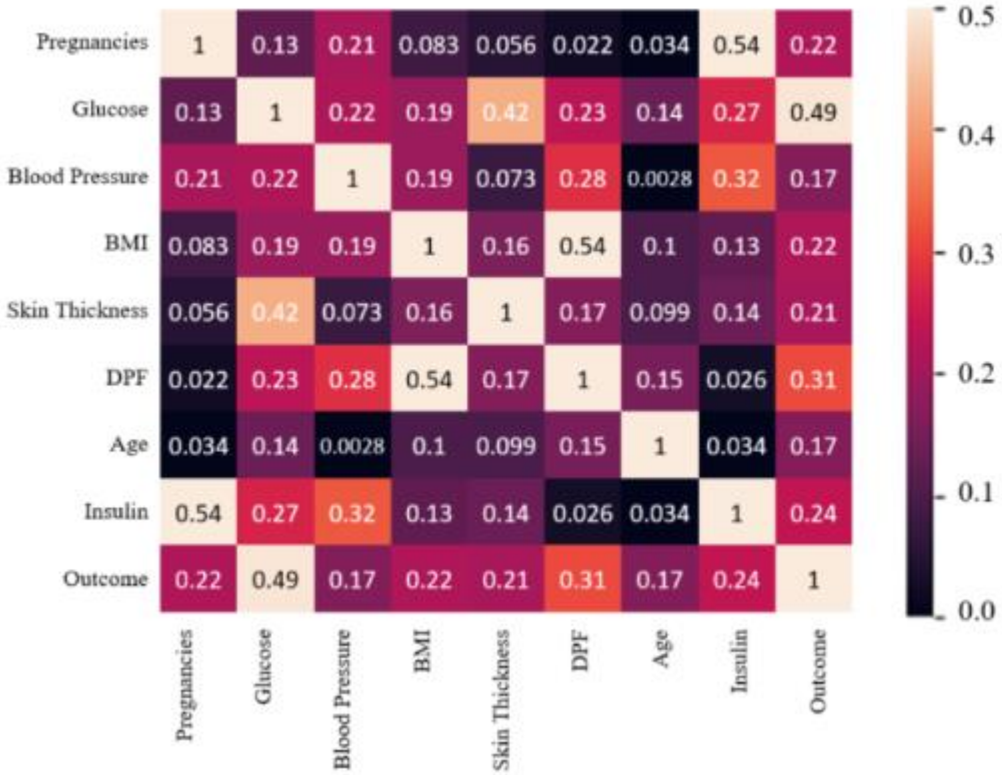
- Glucose shows the strongest correlation (0.49) with diabetes diagnosis
- Diabetes Pedigree Function has the second highest correlation (0.31)
- Other variables show moderate to weak correlations (0.17-0.24)

These correlation values help readers understand which attributes are most predictive of diabetes, informing both the machine learning approach and clinical understanding.

Suggested improvements:

- Include p-values to indicate statistical significance of each correlation
- Sort the attributes by correlation strength for easier interpretation
- Add visual indicators (like + or -) to clarify positive versus negative correlations

Figure 1: Heat Map of Correlation Values



This heat map visualizes the correlation matrix between all variables, using colour intensity to represent correlation strength. The figure shows:

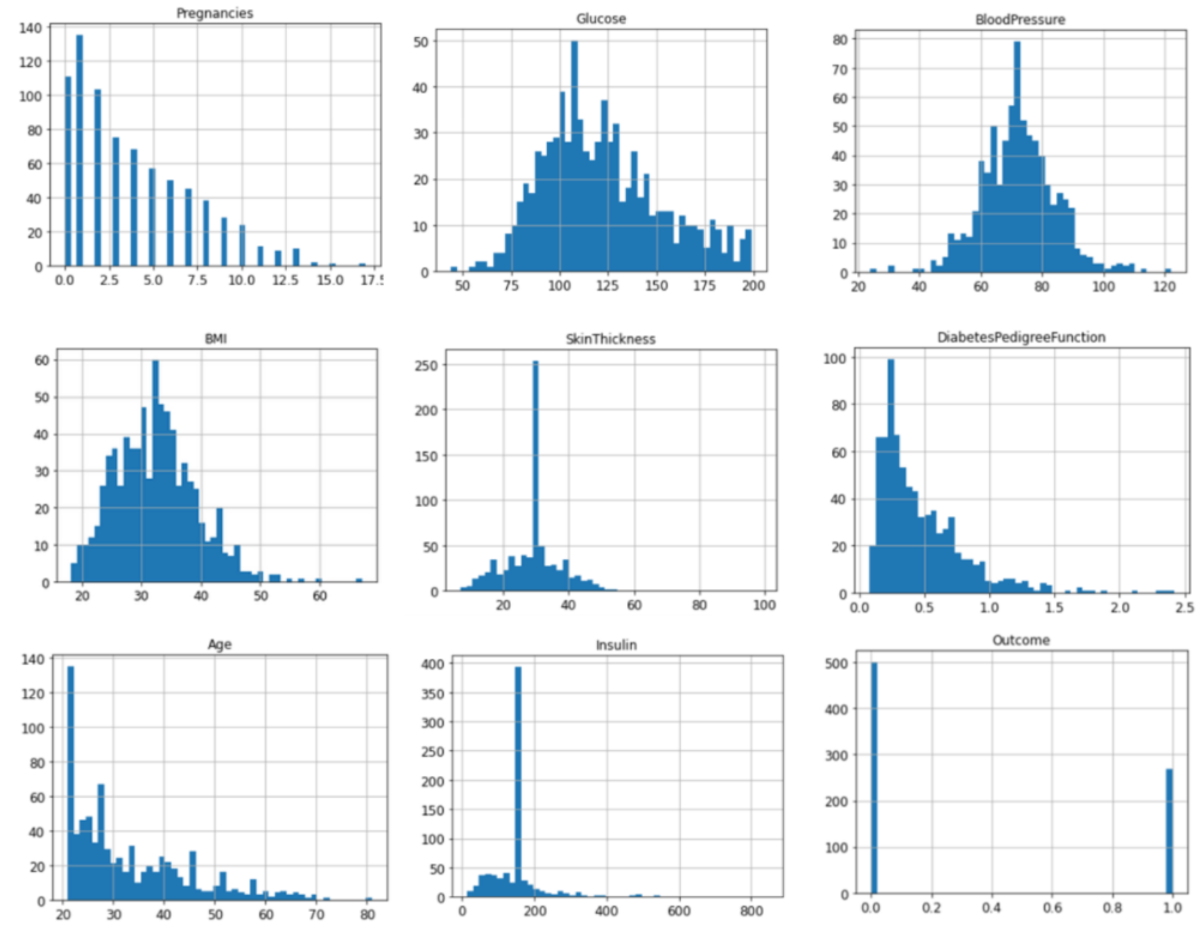
- The diagonal represents perfect self-correlation (1.0)
- Glucose and outcome have the strongest visible positive correlation (matching the 0.49 value from Table 2)
- Age and pregnancy appear to have a moderate positive correlation with each other
- Some attributes show minimal correlation with others (lighter squares)

The heat map effectively condenses a complex matrix of relationships into a visual format, revealing potential multicollinearity issues and reinforcing which variables most strongly predict diabetes.

Suggested improvements:

- Add a colour scale bar to quantify the correlation values
- Label the axes more clearly (the y-axis labels are difficult to read)
- Use a divergent colour scheme (e.g., blue-white-red) to better distinguish positive from negative correlations
- Annotate the strongest correlations with numerical values directly on the heat map

Figure 2: Histograms of the Different Attributes



This figure presents the distribution of each attribute through histograms, showing:

- Most attributes display non-normal distributions
- Some attributes (Glucose, Blood Pressure, BMI) approximate bell curves
- Others (Pregnancies, Insulin, Skin Thickness) show right-skewed distributions
- Age shows multiple peaks, suggesting potential cohort effects
- Outcome shows the class imbalance (more non-diabetic than diabetic cases)

These histograms are crucial for understanding the data distribution and identifying potential preprocessing needs, such as normalization or handling skewed variables.

Suggested improvements:

- Add kernel density plots to smooth the histograms
- Use consistent bin sizes or normalization across attributes
- Add reference lines for mean/median values
- For the Outcome histogram, use descriptive labels ("Diabetic"/"Non-diabetic") rather than just 0/1
- Split histograms by outcome class to show distribution differences between diabetic/non-diabetic groups

Figures 3-8: Confusion Matrices for Different Models

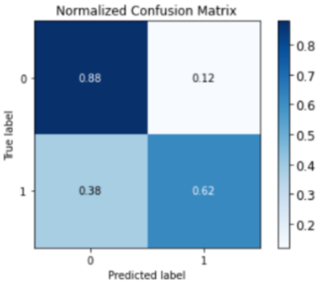


Figure 3: Logistic Regression Confusion Matrix

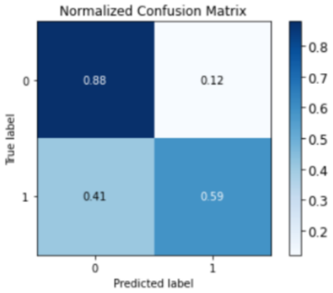


Figure 4: Linear Discriminant Analysis Confusion Matrix

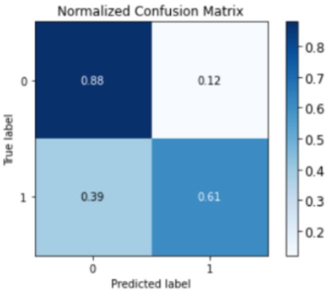


Figure 5: Linear SVM Confusion Matrix

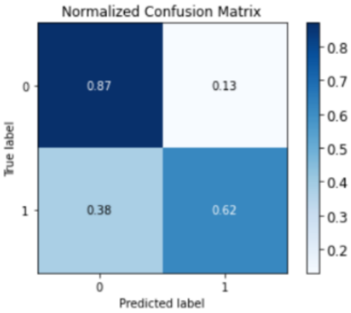


Figure 6: Polynomial SVM Confusion Matrix

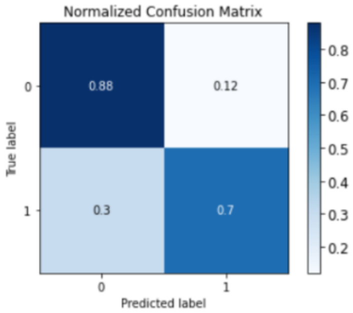


Figure 7: Random Forest Classifier Confusion Matrix

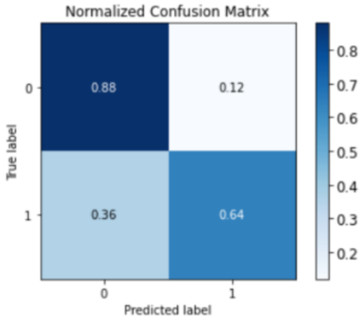


Figure 8: Voting Classifier Confusion Matrix

These six figures display confusion matrices for each classification model tested, showing:

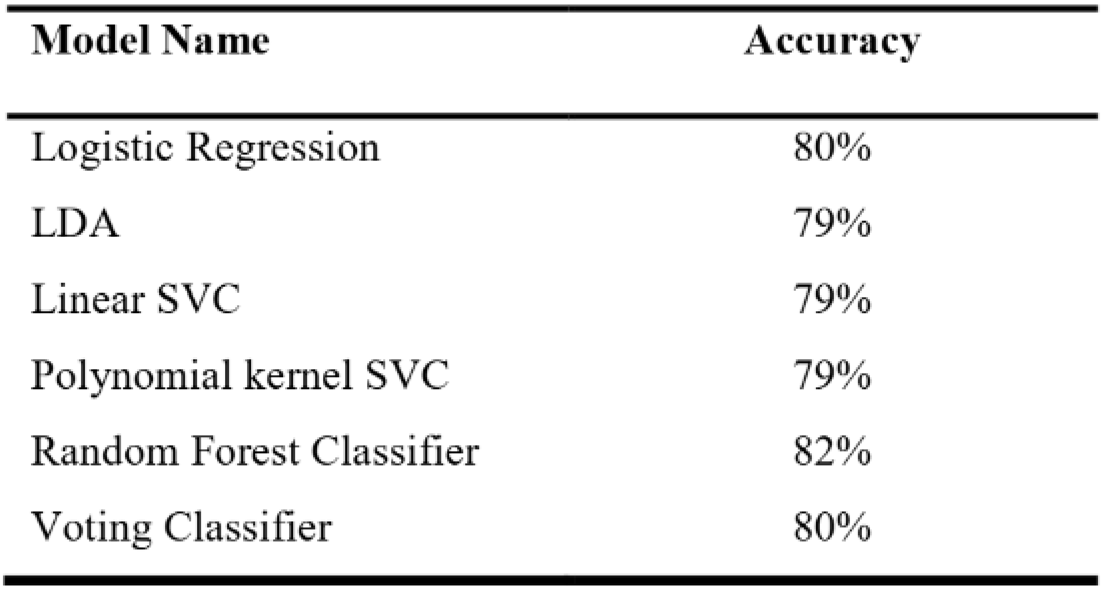
- True positives, false positives, true negatives, and false negatives
- Raw counts rather than percentages
- Similar performance patterns across most models
- Random Forest (Figure 7) shows the highest true positive rate

The confusion matrices provide a deeper understanding of model performance beyond just accuracy, revealing where each model makes errors. This is particularly important for medical diagnostics where false negatives (missing actual diabetes cases) may be more concerning than false positives.

Suggested improvements:

- Add colour intensity or shading to emphasize the diagonal (correct predictions)
- Include derived metrics like precision, recall, and F1-score in each figure
- Standardize the format across all confusion matrices
- Add row and column percentages to better understand error distributions
- Consider combining all matrices into a single figure for easier comparison

Table 3: Accuracy of Trained Models



Model Name	Accuracy
Logistic Regression	80%
LDA	79%
Linear SVC	79%
Polynomial kernel SVC	79%
Random Forest Classifier	82%
Voting Classifier	80%

This final table summarizes the accuracy scores of all six classification models, showing:

- Random Forest Classifier achieved the highest accuracy (82%)
- Logistic Regression and Voting Classifier tied for second place (80%)
- Three models (LDA, Linear SVC, Polynomial kernel SVC) tied at 79%

The table clearly ranks model performance and supports the paper's conclusion that Random Forest was the most effective classifier for this dataset.

Suggested improvements:

- Include additional performance metrics (precision, recall, F1-score, ROC-AUC)
- Add confidence intervals for accuracy scores
- Include computational complexity or training time to assess efficiency
- Highlight the best performer visually
- Add a baseline accuracy (e.g., from always predicting the majority class)

Overall Assessment and Recommendations

The visual elements in this paper effectively communicate both the dataset characteristics and model performance. The authors use appropriate visualization techniques for each purpose: tables for structured information, a heat map for correlation analysis, histograms for distribution analysis, and confusion matrices for performance evaluation.

Recommendations:

1. **Integrated visualizations** that combine multiple metrics (e.g., ROC curves comparing all models)

2. **Feature importance plots** for the Random Forest model to show which attributes contribute most to predictions
3. **Learning curves** to assess how model performance changes with training set size
4. **Calibration plots** to evaluate how well the predicted probabilities match actual outcomes
5. **Visualizations of misclassified instances** to understand systematic errors

These visualizations would strengthen the analysis by providing deeper insights into model behaviour and potential areas for improvement in diabetes prediction.