A Comparison of Machine Learning Algorithms for Diabetes Prediction

Shreya Jadhav (16014223077)
K.J. Somaiya School of Engineering — TY AIDS-B2

Research Paper Overview

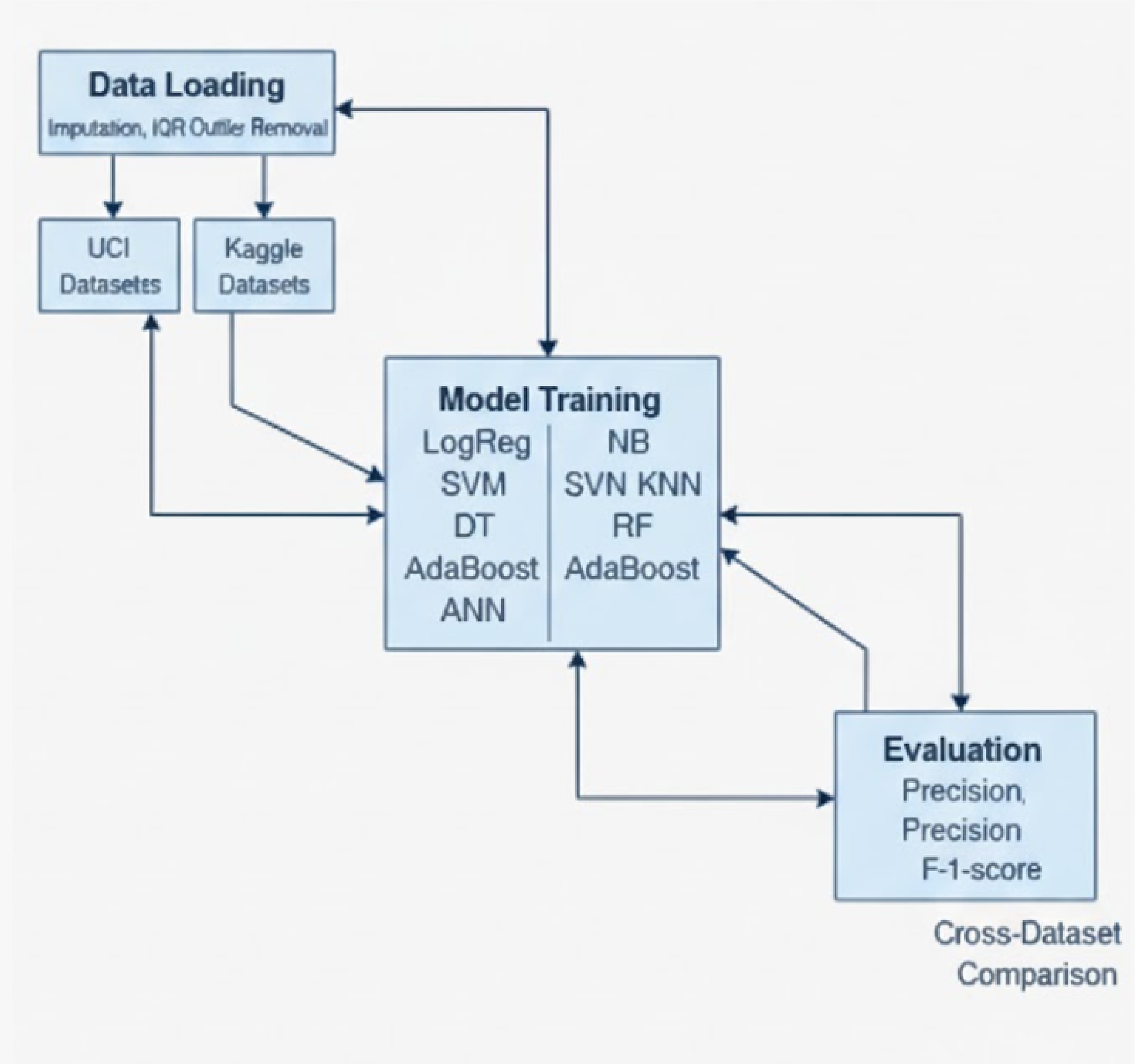
Overview: The referenced study compares classical ML models and ANN for diabetes prediction using the Pima Indians dataset. The paper reports that ANN outperforms other models due to its ability to learn nonlinear feature interactions.

Aim of This Project:

- Reproduce existing methodology.
- Evaluate ML models on Pima and CDC datasets.
- Compare performance between small clinical dataset vs. large real-world dataset.
- Identify the most effective model for diabetes prediction.

System Architecture

- Data Loading (UCI + Kaggle datasets)
- Preprocessing (Imputation, IQR-based outlier removal)
- Feature Selection (Pearson correlation)
- Model Training (LogReg, NB, SVM, KNN, DT, RF, AdaBoost, ANN)
- Evaluation (Accuracy, Precision, Recall, F1-score)
- Cross-Dataset Comparison



Dataset Information

Pima Indians Diabetes Dataset (UCI)

- **Samples:** 768
- **Features:** Glucose, BMI, BP, Insulin, Pregnancies, etc.
- Clinical-only dataset (small + balanced)

CDC BRFSS 2015 Diabetes Dataset (Kaggle)

- **Samples:** 253,680
- **Features:** BMI, Smoking, Activity, HighBP, Income, MentalHealth, etc.
- Real-world dataset (large + imbalanced)

Model Working Process

Step 1: Preprocessing

- Mean imputation for 0-values (Pima)
- Label encoding (CDC)
- IQR-based outlier removal
- MinMax scaling

Step 2: Model Training

- Logistic Regression
- Naïve Bayes
- SVM
- Decision Tree
- Random Forest
- KNN
- AdaBoost
- ANN (2 layers)

Step 3: Evaluation Metrics Accuracy, Precision, Recall, F1-score, Confusion Matrix.

Comparative Results

Phase 1: Pima Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.743590	0.705882	0.444444	0.545455
Naive Bayes	0.730769	0.636364	0.518519	0.571429
SVM	0.756410	0.722222	0.481481	0.577778
Decision Tree	0.615385	0.454545	0.555556	0.500000
Random Forest	0.743590	0.640000	0.592593	0.615385
KNN	0.692308	0.551724	0.592593	0.571429
AdaBoost	0.743590	0.684211	0.481481	0.565217

Performance Summary — Pima Indians Diabetes Dataset

Phase 2: CDC BRFSS 2015 Dataset

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression (Balanced)	0.628149	0.148324	0.795640	0.250036
Naive Bayes	0.917138	0.336449	0.065395	0.109506
SVM (Balanced)	0.694948	0.154393	0.651226	0.249608
Decision Tree (Balanced)	0.628149	0.148324	0.795640	0.250036
Random Forest (Balanced)	0.628149	0.148324	0.795640	0.250036
KNN	0.922092	0.000000	0.000000	0.000000
AdaBoost (Balanced)	0.744622	0.166844	0.570391	0.258171

Performance Summary — CDC Diabetes Indicators Dataset

Best Model (Overall): ANN

- ANN Accuracy on CDC = **93.1%**
- Best handling of nonlinear interactions
- Highest recall for diabetic class

Phase Comparison Summary

Dataset	Best Model	Accuracy	Remarks
Pima Indians (UCI)	SVM	75.6%	Small clinical dataset
CDC BRFSS (Kaggle)	AdaBoost (Balanced)	74.4%	Large + imbalanced

ANN (2-layer) achieved the highest overall accuracy (**93.1%**) on CDC dataset.

Key Observations

- **ANN superior** to all classical ML models.
- CDC dataset enabled strong generalization (size + diversity).
- Balancing improved **Recall** for the diabetic minority class.
- Ensemble methods (AdaBoost, RF) showed stable performance.
- Pima dataset alone is insufficient for large-scale deployment.

Key Insights & Learnings

- **Lifestyle factors** (activity, smoking, mental health) are key risk predictors.
- Normalization/IQR outlier removal improved model stability.
- ANN excels at capturing complex feature interactions.
- Classical ML models failed with imbalanced data.

Interpretation:

- CDC dataset's larger size enhanced model generalization.
- Ensemble methods (AdaBoost, Random Forest) offered stable mid-range per
- Logistic Regression (Balanced) achieved high recall — useful for screening ta

Conclusion

- Successfully implemented and reproduced existing diabetes prediction resear
- Demonstrated scalability of ML algorithms across small and large datasets.
- ANN achieved highest accuracy (93.1%) confirming deep learning's robustnes
- The pipeline can support early detection systems in healthcare applications.

Future Work

Future Work

- Add **Explainability** (SHAP/LIME).
- Deploy interactive Streamlit app prototype.
- Test advanced architectures (Hybrid, Transformers).
- Perform fairness and bias testing.

References

1. Khanam Z., Foo S.Y. (2021), Diabetes Prediction Using ML.
2. Pima Indians Dataset — UCI Repository.
3. CDC BRFSS 2015 Dataset — Kaggle.
4. Mini Project Report, KJSIT (2025).