Chapter 2: Literature Review

2.1 Overview of Machine Learning in Diabetes Prediction

(ML)because of the ability of ML to tackle non-linear relations and massive data volume, the application of ML for early diagnosis of diabetes receives wide popularity. Various types of models from traditionally-used statistical model such as Logistic Regression (LR), to sophisticated algorithms including Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting were involved in the previous researches.

While complex ensemble models often achieve higher accuracy， they usually come with trade-offs in terms of interpretability， training cost， and deployment feasibility— especially in low-resource clinical settings.

2.2 Traditional Machine Learning Models

Traditional ML models offer several advantages， including simplicity， speed， and transparency.The most commonly used models in diabetes prediction include:

(LR)lasso Regression (LR): Known for being computationally fast and simple to interpret.

(DT)decision Tree (DT): Can be used when rules need to be extracted or missing data need to be handled.

(KNN)the KNN algorithm is very simple, but it's easily influenced by the features' scale and imbalanced data.

(NB)version of Bayes (NB): Fast and scalable but assumes feature independence.

-Support Vector Machine(SVM):Effective for high-dimensional data but computationally expensive.

-Random Forest(RF):High accuracy and robustness but less interpretable.

Figure 2. Summary of Traditional ML Models in Literature

Model	Accuracy	Interpretability	Training Speed
Logistic Regression	0.78	High	Fast
Random Forest	0.82	Low	Medium
KNN	0.76	Medium	Slow
SVM	0.75	Medium	Medium
Naive Bayes	0.73	High	Fast
Decision Tree	0.74	High	Fast

Figure 2. Summary of Traditional ML Models in Literature