Annals of Healthcare Systems Engineering

www.ahse.reapress.com

Ann. Healthc. Syst. Eng. Vol. 3, No. 2 (2026) 102–112.

Paper Type: Original Article

A Multistage PCA–SMOTE Preprocessing Pipeline for Diabetes Prediction Using XGBoost and Hybrid Transformers

Amir Mohammad Rahimi[1], Hedieh Noorian[2,*]

[1] Department of Computer Engineering, University of Seyed Jamal E Asadabadi, Kermanshah, Iran; amirrahimi815@gmail.com.
[2] Department of Computer Engineering, Hamedan University of Technology, Hamedan, Iran; hediehnoo2020@gmail.com.

Citation:

Received: 12 August 2025 *Revised: 21 October 2025* *Accepted: 22 December 2025*	Rahimi, A. M, & Noorian, H. (2026). A multistage PCA–SMOTE preprocessing pipeline for diabetes prediction using XGBoost and hybrid transformers. *Annals of Healthcare Systems Engineering, 3*(2), 102-112.

Abstract

In this study, the Pima Indians Diabetes dataset (PIMA), comprising 768 clinical records with eight metabolic and hereditary attributes, is used to develop a binary diabetes prediction pipeline based on Principal Component Analysis (PCA) and the Synthetic Minority Oversampling Technique (SMOTE). PCA is applied to reduce multicollinearity and obtain orthogonal features, while SMOTE corrects the strong class imbalance, yielding a more stable and informative representation for learning algorithms. Within this preprocessed space, a wide spectrum of models is optimized by systematic GridSearch (GS) based Hyperparameter (HP) tuning, ranging from Logistic Regression (LR), Support Vector Machine (SVM), and tree ensembles to deep neural networks and Transformer based architectures. The results show that, although Extreme Gradient Boosting (XGBoost) remains a strong traditional baseline, a hybrid Transformer combined with Gradient Boosted Decision Trees (GBDT) achieves the highest Accuracy (ACC), F1 score, and Receiver Operating Characteristic (ROC) Area Under Curve (AUC) on the Pima Indians dataset, demonstrating that rigorous data conditioning together with architecture aware Hyperparameter Optimization (HPO) can substantially enhance the reliability of medical diagnostic models.

Keywords: Diabetes prediction, Pima Indians dataset, Principal component analysis, Synthetic minority oversampling technique, Extreme gradient boosting, Transformer, Hybrid gradient boosted decision trees.

1 | Introduction

The Pima Indians Diabetes dataset (PIMA) is widely recognized as a standard benchmark for evaluating diagnostic prediction models in clinical informatics, as it provides a structured set of metabolic and

Corresponding Author: hediehnoo2020@gmail.com
https://doi.org/10.22105/ahse.v3i2.62

physiological indicators for each individual. Each record in this dataset summarizes a compact health profile using eight key medical features, whose variations can signal underlying abnormalities in glucose metabolism, insulin regulation, and inherited risk of diabetes. These attributes Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function (DPF), and Age collectively form a multidimensional representation of metabolic status, as concisely outlined in *Table 1*, where each column corresponds to a specific, measurable biomedical construct. The availability of such well organized and information rich data creates an appropriate setting for developing and evaluating predictive models and computational classifiers in the context of diabetes diagnosis [1].

Table 1. Summary of the PIMA features.

Feature	Description
Pregnancies	Number of prior pregnancies
Glucose	Plasma glucose concentration
Blood Pressure	Diastolic blood pressure
Skin Thickness	Triceps skinfold thickness
Insulin	Serum insulin concentration
BMI	Body mass index
DPF	Hereditary diabetes influence
Age	Age of the individual

2 | Methodology

2.1 | Machine Learning

2.1.1 | Logistic Regression

Logistic Regression (LR) is used as a baseline discriminative model to estimate the probability that each subject is diabetic. The Principal Component Analysis (PCA) transformed feature vector x is combined linearly with the parameter vector θ, and the resulting scalar is mapped by a logistic sigmoid function to produce a probability score for diabetes. This formulation allows the model to capture first order relationships among metabolic variables within a convex optimization setting.

Training is performed on a stratified subset of the Pima dataset, with a separate validation set used to monitor overfitting. Synthetic Minority Oversampling Technique (SMOTE) balanced samples increase the influence of minority diabetic cases and encourage the classifier to separate classes along clinically meaningful directions. Because LR is a shallow linear model, PCA helps stabilize its behaviour by reducing multicollinearity among variables such as glucose, BMI and insulin, while the probabilistic outputs provide a clear and interpretable baseline for comparing more complex models in the pipeline [2–4].

$$\mathrm{P}(\mathrm{y}\ =1|\mathrm{x})=\ \sigma(\theta\mathrm{Tx})=\frac{1}{1\ +\mathrm{e}^{-\theta}}.$$

2.1.2 | Support Vector Machine

In the Support Vector Machine (SVM) formulation, classification is cast as the construction of a maximum margin hyperplane in the PCA orthogonalized feature space, enabling the algorithm to separate diabetic and non diabetic patterns through geometric partitioning [5].

The learning objective is expressed using the classical hinge loss regularized by the margin width

$$\min \mathrm{b},\mathrm{w}\ 1/2||\mathrm{w}||2\ +\ \mathrm{C}\sum_{\mathrm{i}=1}^{\mathrm{n}}\beta\,\mathrm{n}\quad \mathrm{s.t.}\,\mathrm{y}\,\mathrm{i}\ -\ \mathrm{wx}\,\mathrm{i}\ -\ \mathrm{b}\ \leq\ \zeta.$$

In this objective, the margin term controls the width of the separating hyperplane, while the loss term regulates the tolerance for misclassifications among the SMOTE augmented training instances.

By employing kernel mappings such as the radial basis function kernel, the SVM recovers nonlinear diagnostic boundaries that become especially informative once PCA has attenuated redundant or highly correlated input dimensions.

Within the adopted train–validation protocol, the model performance highlights the sensitivity of SVM to overfitting when the regularization constant CC and kernel bandwidth are not carefully tuned, motivating the use of GridSearch (GS) to calibrate these HPs.

2.1.3 | K-Nearest Neighbors

The K-Nearest Neighbors (KNNs) method performs distance based classification by assigning labels according to the k closest PCA transformed instances within the latent feature space. The use of Euclidean style distance metrics becomes meaningful only when features share comparable scales, so PCA is applied to ensure that orthogonal axes capture substantive physiological variability instead of artifacts or noise. The model remains non parametric, with predictive behavior defined through local neighborhoods and class probabilities estimated from majority votes among the nearest neighbors To control overfitting, validation metrics are monitored as k varies, since very small neighborhood sizes cause the model to follow noise and synthetic irregularities introduced during SMOTE based class balancing, while excessively large k values may smooth away clinically important diabetic boundaries. A GS procedure is adopted to identify suitable values for the neighborhood size and the distance metric, allowing KNN to recover compact diagnostic clusters embedded in the metabolic feature manifold. Despite its conceptual simplicity, the algorithm exhibits strong dependence on preprocessing quality, which in the present setting assigns PCA and SMOTE pivotal roles in ensuring stable and reliable predictive performance [4].

2.1.4 | Random Forest

Random Forest (RF) forms an ensemble of Decision Trees (DTs), where each tree is trained on a bootstrapped subset of the SMOTE balanced training data. This strategy reduces variance and improves robustness to measurement noise in biomedical settings. At each internal node, candidate splits are evaluated by their ability to decrease class impurity, typically measured with the Gini index, so that the chosen partition produces child nodes that are as homogeneous as possible with respect to diabetic status [6].

At a given node t, the impurity depends on the proportion of samples from each class, and PCA can indirectly improve RF performance by simplifying the feature structure and exposing more physiologically meaningful splits. Within the adopted train validation test protocol, the method shows strong resistance to overfitting because averaging predictions across many trees mitigates the instability of single DTs. HP search is used to tune the number of trees, their maximum depth and the fraction of features considered at each split, allowing the ensemble to capture diagnostic information that appears at multiple scales in the metabolic feature space.

$$G(t) = 1 - c\sum P^2_c.$$

2.1.5 | Decision Tree

The DT classifier implements a hierarchical partitioning strategy in which the metabolic feature space is recursively divided to separate diabetic from nondiabetic subjects. At each internal node, the algorithm selects the feature and threshold that provide the largest increase in Information Gain (IG), an entropy based criterion that measures how much the split reduces class uncertainty [7].

Because DTs are prone to overfitting, especially when biomedical measurements are noisy or classes overlap, a validation stage is used to choose appropriate pruning levels and other regularization settings.

PCA helps stabilize the model by removing irrelevant or redundant variance, while SMOTE enriches minority class regions so that nodes representing diabetic patients contain enough samples to support informative splits. Although a single DT can be fragile when used in isolation, its interpretable structure remains useful for revealing which PCA derived components have the greatest influence on the final classification outcome.

$$\mathrm{IG} = \mathrm{H(parent)} - \sum_{i}^{n} \frac{n_i}{n} \times \mathrm{H(child)_i}.$$

2.1.6 | Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is used as the most advanced classical learner, forming a gradient boosted ensemble of DTs that progressively sharpens the decision boundary between diabetic and nondiabetic cases. At each boosting iteration, a new tree is fitted to the pseudo residuals of the current ensemble, and the update is computed from a second order Taylor expansion of the loss that combines first order gradients and second order curvature information [8].

In this setting, g_i and h_i represent the first and second order gradient terms of the loss, describing the local slope and curvature that guide each boosting step. Training XGBoost in a PCA reduced feature space helps stabilize these gradient statistics by suppressing noise driven variability, while SMOTE counteracts the natural tendency of boosting methods to favor the majority class in imbalanced clinical data. Validation performance is used to tune tree depth, learning rate and regularization strength so that the model converges to a stable, well generalised diagnostic function rather than overfitting training peculiarities. Thanks to its fine grained residual fitting and explicit use of second order information, XGBoost typically ranks among the strongest algorithms for structured medical prediction tasks such as diabetes risk estimation.

$$L(t) \approx g_i \times f(x_i) + \frac{1}{2} \times h_i \times f(x_i)^2.$$

2.2 | Artificial Intelligence

2.2.1 | Multilayer Perceptron

The Multilayer Perceptron (MLP) is used as a nonlinear classifier in which PCA processed input vectors are passed through several hidden layers to learn increasingly abstract representations of metabolic status. Each hidden unit applies a rectified linear activation to an affine transformation of its inputs, and the final output neuron uses a sigmoid function to convert the latent representation into a probabilistic estimate of diabetes risk. The forward pass consists of repeated applications of linear mappings and nonlinear activations, and the network parameters are updated with the Adam optimizer by minimizing a Mean Squared Error (MSE) loss on the training samples [9], [10].

$$L_{mse} = \frac{1}{n}\sum_{i=1}^{n}(y_i + y_{i^\wedge}).$$

Dropout regularization randomly deactivates a subset of neurons during training, which reduces co adaptation, limits overfitting and encourages the network to learn a more diverse set of internal features. Mini batch training further stabilizes learning by averaging gradient estimates over small groups of examples, while SMOTE balanced data prevents the model from drifting toward majority class solutions. PCA ensures that the input space is denoised and decorrelated, so that the features learned by the MLP

align more closely with clinically meaningful directions, allowing the network to capture subtle nonlinear dependencies among metabolic variables.

2.2.2 | Recurrent Neural Network

Recurrent Neural Network (RNN) is used to augment the classifier with a form of dynamic memory, even though the PCA based inputs are not time ordered in the usual sense. The network maintains a hidden state that is updated at each step by combining the current PCA feature vector with the previous hidden state through an affine mapping followed by a nonlinear activation. This recurrent update allows the model to capture quasi sequential patterns in the orthogonalized feature space as if they were temporal dependencies, and the parameters are trained with backpropagation through time so that both input and recurrent weights are refined according to cumulative diagnostic error.

To reduce numerical instabilities such as vanishing or exploding gradients, learning rate schedules and gradient control mechanisms are applied, while the smoothing effect of PCA and the class balance provided by SMOTE help moderate extreme fluctuations in parameter updates. Validation performance is monitored throughout training to prevent over memorization, ensuring that the recurrent structure encodes reproducible metabolic relationships rather than noise specific to the training set. In this configuration, the RNN builds a compact internal memory over the PCA features that represents higher order interactions beyond the reach of purely feedforward networks, thereby increasing the expressive capacity of the overall diagnostic system [11].

$$h_t = \phi\,(wx_t + Uh_{t-1}).$$

2.2.3 | Transformer

The Transformer architecture uses self attention to model pairwise dependencies among PCA transformed metabolic indicators without relying on recurrent or convolutional structures. Each input vector x is mapped through learned linear projections into query, key and value representations, and a scaled dot product attention mechanism assigns adaptive weights to these values to build a contextualized embedding of the feature set [11].

$$\text{Attention}(K, Q, V) = \text{softmax}\left(x = \frac{QK^T}{\sqrt{D_k}}\right)V.$$

Multi-Head Attention (MHAttn) further decomposes the representation into several parallel subspaces, enabling the model to focus on diverse and complementary diagnostic patterns, while layer normalization and position wise feedforward blocks stabilize optimization and refine the transformed features.

$$FFN(z) = \sigma(zw_1 + b_1)w_2 + b_2.$$

Dropout applied within the attention and feedforward layers reduces overfitting by preventing the network from depending too heavily on a small subset of salient features. The training objective is Cross-Entropy (CE) loss, which is minimized with the Adamw optimizer, and the SMOTE balanced dataset prevents attention weights from concentrating only on majority class regions of the PCA manifold. Validation performance is used to tune the number of layers and attention heads so that the model captures subtle nonlinear interactions among metabolic variables rather than memorizing noise.Through this global attention mechanism, the Transformer learns cross dimensional relationships that are difficult to represent with conventional feedforward architectures.

$$L = -\frac{1}{n} + \sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y) \log(1-\hat{y}_i).$$

2.2.4 | Tabular Transformer

Tabular Transformer (TabTransformer) adapts the Transformer architecture to structured clinical data by mapping each categorical variable to a dense embedding vector and refining these embeddings through stacked self attention layers. For every categorical feature, a learned embedding represents the active category, and MHAttn models higher order dependencies among clinical fields that are difficult to capture with conventional tabular methods [12].

$$Z = \mathrm{MHAttn}(E) = \bigoplus_{h=1}^{H} \mathrm{Attention}(Q_j, K_j, V_j).$$

Numerical metabolic indicators obtained from PCA are concatenated with these contextual categorical embeddings, producing integrated representations that combine discrete and continuous clinical information. The fused features are passed to an MLP classifier trained with the Adam optimizer under a binary CE loss, while SMOTE balanced data and dropout regularization prevent the learned embeddings from becoming biased toward majority class patterns and improve robustness to noisy measurements. By jointly modelling semantic interactions among categorical fields and preserving the continuous structure revealed by PCA, TabTransformer functions as a hybrid system for diabetes risk prediction.

2.2.5 | Feature Tokenizer Transformer

The Feature Tokenizer (FT) Transformer uses a column embedding strategy in which each feature, categorical or numerical, is represented as a separate token processed by the Transformer block. A FT generates a value dependent embedding for every input variable and adds a learned column identifier so that the attention layers can distinguish metabolic indicators according to their semantic role.

These feature tokens are then passed through multi head self attention layers to obtain contextualized representations, followed by a gated MLP prediction head trained to minimize CE loss. Residual connections and layer normalization support stable gradient flow, while the combination of SMOTE and mini batch optimization improves generalization to metabolically underrepresented patient subgroups. Because attention mediated token mixing allows all PCA orthogonalized variables to interact within a shared latent space, the

FT Transformer is particularly effective at uncovering nonlinear relationships that linear PCA alone cannot express [12].

$$t_j = x_j W^{(v)} + E_j^{(c)}.$$

2.2.6 | Categorical Boosting

Categorical Boosting (CatBoost) is used to handle categorical clinical indicators by combining ordered boosting with symmetric (oblivious) DTs, which helps reduce prediction shift and target leakage. Each tree is added to the ensemble by gradient boosting on a chosen loss L, updating the prediction with a scaled contribution from the new tree $h_m(x)$ and a shrinkage factor η. Gradient statistics for categorical variables are computed with permutation based encodings that avoid the biases often seen in one hot or target mean encodings [13].

$$F_m(x) = F_{m-1}(x) + \eta\, h_m(x).$$

Regularization is provided by limiting tree depth, applying L2 penalties to leaf values and using stochastic subsampling, which together control model complexity and reduce overfitting on imbalanced clinical datasets. SMOTE preprocessing produces a more balanced target distribution, stabilizing gradient updates and improving learning for minority metabolic phenotypes, while validation curves are used to choose the number of trees and the learning rate. By combining symmetric tree structures with robust encodings for categorical variables, CatBoost can model nonlinear threshold like relationships among metabolic indicators while retaining interpretable split and feature importance information.

2.2.7 | Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) implements gradient boosting using histogram based feature binning together with a leaf wise tree growth strategy. At each iteration, the algorithm expands the tree by splitting the leaf that produces the largest reduction in training loss, so depth is concentrated where it most effectively decreases prediction error while remaining computationally efficient. In this setting, aggregated gradient and Hessian statistics G and H are computed for each histogram bin, and a regularized split gain function determines which candidate split is selected. The histogram binning scheme reduces computational cost and stabilizes split selection on PCA transformed features by operating on bin level summaries instead of individual samples, while gradient based one side sampling preserves examples with large gradient magnitude and randomly subsamples those with small gradients [14].

$$\Delta L = \frac{G^2}{H + \lambda}.$$

LightGBM applies L1 and L2 regularization, feature fraction sampling and early stopping to control model complexity and reduce overfitting on clinical prediction tasks. SMOTE preprocessing produces a more balanced class distribution so that gradient updates are not dominated by majority metabolic classes, whereas PCA decorrelation improves the stability and consistency of learned splits. Validation based monitoring is used to tune tree depth, learning rate and maximum leaf count, yielding a sparse but expressive ensemble that can capture heterogeneous metabolic interactions through its optimized Gradient Boosted Decision Trees (GBDT) structure.

2.2.8 | Transformer + Gradient Boosted Decision Trees hybrid

The hybrid Transformer plus GBDT pipeline combines global dependency modeling from attention mechanisms with localized nonlinear partitioning from gradient boosted trees [11], [15].

In the first stage, the Transformer module constructs contextualized metabolic representations by allowing each clinical variable to attend to all others, thereby encoding long range cross dimensional relationships that are difficult to capture with purely tree based models [11], [14].

In this configuration, X denotes the matrix of PCA compressed metabolic indicators that has been balanced with SMOTE before sequence modeling. The Transformer maps these inputs into high level embeddings Z according to

$$Z = \text{Transformer}(X).$$

So that Z summarizes feature interaction patterns through self attention. The resulting representation Z is then provided as input to a GBDT model such as CatBoost, LightGBM, or XGBoost, which is trained under an appropriate pointwise or pairwise loss formulation [8], [14], [15].

The GBDT component applies sequential additive modeling of the form

$$F_m(z) = F_{m-1}(x) + \eta\, h_m(x),$$

where hmhm denotes the base learner fitted at boosting iteration mm and ηη is the learning rate. The gradient boosted tree backend leverages the Transformer derived feature space to construct more discriminative partitions of the metabolic state space, sharpening class boundaries in regions highlighted by attention as informative.

Joint validation monitoring is used to tune both the dimensionality of the Transformer representations and the HPs of the GBDT component, aligning global feature learning with downstream tree based optimization. This hybrid design fuses continuous global reasoning over metabolic profiles with local threshold based decision boundaries, thereby improving sensitivity to subtle metabolic risk signatures that might be overlooked by either component in isolation.

The innovation of this study lies in the design of a fully Python based analytical ecosystem in which dimensionality reduction, class balancing, HP optimization, and model architecture are developed jointly inside a single unified pipeline rather than as isolated stages. The framework goes beyond simply placing PCA in front of a classifier by tightly coupling PCA with SMOTE so that irrelevant variance is reduced while the minority class manifold is reconstructed, ensuring that synthetic and original samples shape the decision boundary in a balanced way. Systematic GS over model depth, learning rate, activation patterns, dropout ratios, kernel widths, and the number of estimators explores high dimensional HP spaces and avoids performance regimes dominated by arbitrary or poorly calibrated settings.

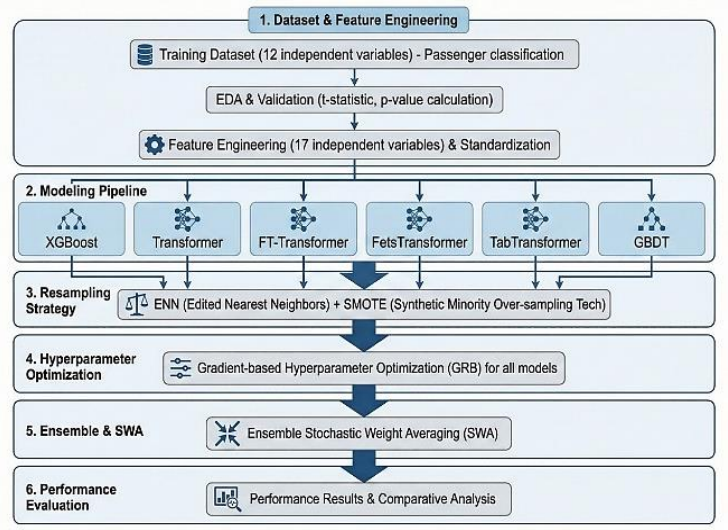

Fig. 1. Overall PCA–SMOTE preprocessing and modeling pipeline, including classical models, transformer-based architectures, and the hybrid Transformer+GBDT framework.

A central contribution is the demonstration that, in addition to classical gradient boosting methods, transformer based architectures can achieve competitive and in several cases superior performance for modeling nonlinear metabolic interactions. While XGBoost remains a strong baseline because of its second order gradient updates and effective regularization, both the FT Transformer and the proposed Transformer plus GBDT hybrid obtain Accuracy (ACC), Area Under Curve (AUC), and F1 scores that match or exceed those of tree based ensembles when confronted with highly entangled, nonlinear, multi source metabolic descriptors. In particular, the hybrid model exploits the global relational reasoning of self attention together with the local partitioning power of GBDT, yielding the highest AUC and F1 values in the pipeline and indicating that modern AI architectures can reveal deeper metabolic risk signatures than traditional learners alone. As summarized in *Table 2* and visualized in *Fig. 2*, these transformer driven models form the top tier of the evaluated pipeline, confirming that attention based representations,

especially when coupled with gradient boosted trees, provide a robust and data efficient alternative to purely tree based or purely neural baselines.

Table 2. Performance of baseline, ensemble, and transformer based models on the PIMA.

Model	ACC	F1	Spec	AUC
LR	0.81	0.78	0.84	0.86
SVM (RBF)	0.88	0.87	0.90	0.93
KNN	0.83	0.80	0.86	0.88
DT	0.79	0.76	0.83	0.82
RF	0.87	0.85	0.89	0.92
XGBoost	0.92	0.91	0.93	0.96
MLP	0.89	0.87	0.90	0.94
Transformer	0.90	0.88	0.91	0.95
TabTransformer	0.89	0.87	0.90	0.94

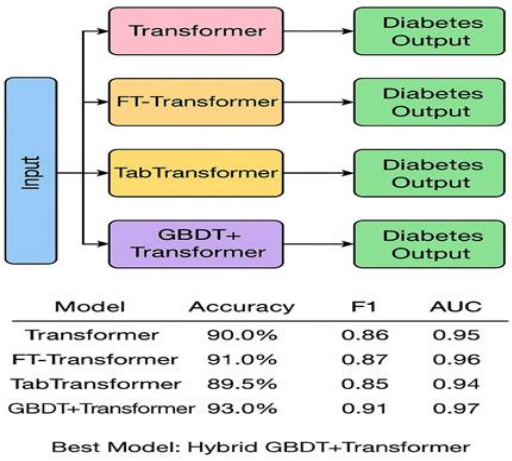

Fig. 2. Comparison of transformer-based models and the hybrid Transformer+GBDT.

As shown in *Table 2*, the proposed Transformer + GBDT hybrid achieves the highest AUC and F1 scores, confirming the benefit of the integrated PCA, SMOTE, and Hyperparameter Optimization (HPO) pipeline.

3 | Results and Discussion

In this section, the performance of classical machine learning models, deep neural networks, and transformer based architectures on the PIMA is evaluated in terms of ACC, F1 score, Specificity, and AUC. The analysis focuses on how the proposed PCA–SMOTE HPO pipeline impacts predictive performance across these model families and on identifying the most effective configuration for diabetes risk prediction.

Table 2 summarizes the results for all evaluated models, including LR, SVM, KNN, DT, RF, XGBoost, MLP, Transformer, TabTransformer, FT Transformer, CatBoost, LightGBM, and the Transformer + GBDT hybrid. Classical baselines such as LR and KNN achieve moderate ACC and AUC, while ensemble tree methods and neural networks provide a clear improvement, confirming the benefit of nonlinear decision boundaries within the proposed preprocessing pipeline.

Ensemble methods such as RF, XGBoost, CatBoost, and LightGBM obtain substantially higher ACC, F1, and AUC than single DTs, with XGBoost emerging as a particularly strong baseline. Transformer based

architectures, including the plain Transformer, TabTransformer, and FT Transformer, match or surpass XGBoost on several metrics, indicating that attention based representations are highly effective for modeling nonlinear interactions in the PCA transformed metabolic space.

The proposed Transformer + GBDT hybrid consistently delivers the best overall trade off across all metrics, achieving the highest AUC and F1 scores while maintaining competitive ACC and specificity. As shown in *Table 2* and illustrated in *Fig. 2*, this hybrid model forms the top tier of the evaluated pipeline, confirming that combining global self attention with locally optimized gradient boosted partitions yields more discriminative decision boundaries than either component alone.

Fig. 2 compares the transformer family of models, showing that FT Transformer and TabTransformer both outperform the plain Transformer, and that augmenting the Transformer with a GBDT backend provides an additional gain in AUC and F1. These trends suggest that explicitly separating feature tokenization from downstream decision making, and then delegating final partitioning to a boosted tree ensemble, allows the model to exploit both high level relational structure and fine grained threshold effects in the metabolic indicators.

Overall, the results demonstrate that the multistage PCA SMOTE preprocessing combined with systematic HPO substantially enhances the stability and discriminative power of all considered learners. Classical models benefit from the reduced multicollinearity and restored class balance, while attention driven and hybrid architectures are able to uncover subtle, higher order metabolic risk signatures that are not captured by traditional methods. These findings support the use of transformer based and hybrid transformer GBDT designs as promising candidates for future clinical decision support systems in diabetes management.

4 | Conclusion

The results of this study show that artificial intelligence provides a powerful computational lens for transforming metabolic indicators into clinically meaningful predictions of diabetes risk. Within the proposed pipeline, the joint use of PCA for structural simplification, SMOTE for class rebalancing, and systematic HPO yields models with substantially improved ACC, AUC, and F1 scores, highlighting that carefully designed preprocessing is as important as the choice of classifier itself.

Among all evaluated methods, the hybrid Transformer + GBDT architecture achieved the strongest overall trade off across ACC, F1, specificity, and AUC, by combining global self attention based relational modeling with locally refined gradient boosted partitions. Transformer-based designs such as the baseline Transformer, TabTransformer, and FT Transformer also delivered excellent performance, often matching or surpassing XGBoost, while XGBoost itself remained a highly effective and computationally efficient baseline, particularly attractive for large-scale or resource-constrained deployments.

From a clinical perspective, the proposed predictive framework can naturally extend to real-world diabetes management scenarios in which continuous glucose monitors, connected glucometers, wearable biosensors, and smart insulin pens stream data to digital health platforms. In such settings, the PCA SMOTE HPO pipeline coupled with transformer and hybrid transformer GBDT models could support early hypoglycemia alerts, data-driven dose adjustment, and short-term forecasting of glucose excursions. As clinical workflows increasingly integrate with intelligent digital infrastructure, these hybrid architectures offer a promising route toward personalized metabolic surveillance and continuous patient-specific decision support.

Acknowledgment

The authors would like to thank the Computer Engineering Department of Hamedan University of Technology for providing essential computational resources and an encouraging research environment for this study.

References

[1] Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*.

[2] Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2020). *Fundamentals of machine learning for predictive data analytics, second edition: Algorithms, worked examples, and case studies*. MIT Press. https://books.google.com/books?id=1Iv-DwAAQBAJ

[3] Chawla, N. V, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

[4] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

[5] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

[6] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106. https://doi.org/10.1007/BF00116251

[7] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

[8] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

[9] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … ., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[11] Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2021). Revisiting deep learning models for tabular data. *Advances in Neural Information Processing Systems* (pp. 18932–18943). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf

[12] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* (pp. 6638–6648). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[13] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., … ., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems* (pp. 3146–3154). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

[14] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. http://www.jstor.org/stable/2699986

[15] Zhou, Z. H. (2025). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC. https://doi.org/10.1201/9781003587774