Early-Stage Diabetes Risk Prediction using Explainable and Uncertainty-Aware Deep Learning: A Multi-Domain Framework with Selective Prediction and Polynomial Feature Stacking

[1]Dr. Bimal Dutta, [2]Akram Hossain*
[1]Professor, [2]M.Tech Student
[1]Department of Computer Science and Engineering, [2]Department of Computer Science and Engineering
[1, 2]Narula Institute of Technology Agarpara, Kolkata, West Bengal, India

***Abstract* - Disease T2DM is an undiagnosed, fast-rising disease prevalent in adults with a worldwide incidence of 537 million individuals; thus, the need to identify a method to predict this disease early and by non-invasive means is a pressing challenge. Nonetheless, current machine learning models have limited generalizability, do not estimate uncertainty, and cannot be applied clinically, which are important research gaps. The research paper is aimed at filling in the gap of uncertain and reliable predicted diabetes systems, and extrapolating on heterogeneous data sets.**

This study aims to create a robust and uncertainty-sensitive diabetes prediction model applicable to heterogeneous data and aiding clinical decision-making. The suggested multi-domain framework integrates the data harmonisation realized through MICE with the expansion of the number of features to a polysomnoid (844 features), and uncertainty-aware deep learning and calibrated stacking. The system incorporates uncertainty estimation, calibration, and selective prediction to enhance the reliability of clinical decision-making. Uncertainty on a per-patient basis is offered by a residual neural network with Monte Carlo dropout, and selective prediction is yet another method that helps to increase the trustworthiness of the model.

The results of the experiment show a great improvement, with an AUC of 0.908 (Random Forest = 0.797 and Logistic Regression = 0.735) and high calibration (ECE = 0.031, Brier = 0.128). The given system is intended to be used in real-life screening situations when the reliability of prediction and interpretability are crucial factors. In sum, the current piece of work offers a clinically implementable framework that can be used to promote the screening of diabetes and informed decision-making.

Keywords: **diabetes prediction, Monte Carlo Dropout, selective prediction, MICE imputation, polynomial features, stacking ensemble, SHAP, uncertainty quantification, domain adversarial learning.**

I. INTRODUCTION

Type 2 diabetes mellitus is one of the greatest health problems among the world population, with an estimated 537 million individuals having diabetes and a projected figure of 783 million by 2021. Although its incidence is on the rise, the disease may remain silent so that patients can have normal fasting plasma glucose and HbA1c levels over a long period of time, and insulin resistance occurs silently. As a result, most people are diagnosed when complications have already developed, including neuropathy, retinopathy, or cardiovascular disease. The issue of late diagnosis underscores the necessity of a timely, non-invasive, and scalable screening instrument that can help to detect high-risk patients before it is too late to take action. Machine learning (ML) has demonstrated significant potential in diabetes prediction in recent years, providing automated assessment of the risks of the disease, which is based on clinical data and demographic characteristics. Nevertheless, even with an extensive literature, the application of these models to real clinical practice has been low because there are significant gaps in their generalisability, quantification of uncertainty, and their readiness to be deployed in practice.

One of the key limitations of the current literature is that most research has not been substantively validated with various datasets. Most models are created and experimented on one dataset, the Pima Indians Diabetes Dataset, but are controlled. Although these models typically achieve high-performance (e.g., AUC > 0.90), studies often fail to be accurate in other contexts involving more

heterogeneous, real-world populations with different data distributions, measurement protocols, and feature availability. Recent research has tried cross-dataset validation, although not many have explicitly done systematic research on domain shift with systematic evaluation like leave-one-domain-out (LODO) or domain-adversarial training. Moreover, in the existing methods, the estimation of uncertainty is not considered much. The majority of classifiers only give point predictions and do not specify the confidence or reliability of the prediction, which limits their clinical use in situations where risk stratification and care in cases of uncertainty are needed. Moreover, the current frameworks seldom incorporate interpretability, calibration, and selective forecasting into one system, providing clinicians with no clear or practical information.

To overcome these shortcomings, in this study, I plan to design a holistic, clinically implementable machine learning system to predict early T2DM by placing greater emphasis on generalisability, awareness of uncertainty, and interpretability. In particular, the goals are: (i) to build a harmonised multi-domain dataset that measures the heterogeneity of populations, (ii) to refine feature representations by being able to expand their polynomials to capture complex biomarker interactions, (iii) to build a deep-learning model that can estimate epistemic uncertainty by using Monte Carlo Dropout, (iv) to build a selective prediction mechanism to increase reliability by not making uncertain predictions The originality of the work, in contrast to earlier research that investigates the enhancement of individual models, is that data harmonisation, uncertainty quantification, interpretability (through SHAP) and cross-domain validation are incorporated at the system level in a single integrated pipeline.

Existing machine learning models for diabetes prediction often focus on accuracy but fail to address uncertainty, calibration, and cross-domain generalization.
Notably, this framework is clinically deployment-oriented. With every prediction, one can see a score above the probability, an estimate of uncertainty, and a feature-level explanation, which allows clinicians to make a judgment about the intervention or additional tests. The selective prediction component also matches the real-world working processes since it highlights high-uncertainty cases to be further evaluated instead of imposing potentially invalid forecasts. The rest of this paper is structured as shown: (2) related work reviews and identifies limitations of current ML-based and diabetes prediction models, (3) datasets and the harmonisation pipeline proposed by MICE, (4) feature engineering and model architectures, (5) experimental results, cross-domain validation and uncertainty analysis, (6) clinical implications and deployment considerations, and (7) conclude the study with future research directions. The objectives are below:

- To develop a multi-domain diabetes prediction framework by integrating heterogeneous datasets using MICE harmonisation for improved generalisability.
- To enhance predictive performance by modelling nonlinear biomarker interactions through polynomial feature expansion and stacking ensemble learning.
- To incorporate uncertainty-aware deep learning using Monte Carlo Dropout and selective prediction to improve reliability in clinical decision-making.
- To provide interpretable and clinically deployable predictions using SHAP-based explanations, calibrated probabilities, and validated cross-domain performance.
- To develop a reliable and uncertainty-aware diabetes prediction framework that generalizes across heterogeneous datasets and supports clinical decision-making

II. RELATED WORK

A. Machine Learning for Diabetes Prediction

Classical ML benchmarks on single-cohort diabetes datasets are well-established. Sisodia and Sisodia [5] compared Naive Bayes, Decision Tree, and SVM on the Pima dataset, achiev-ing 76.3 % accuracy. Zou et al. [6] showed Random Forest superiority on glycaemic outcome prediction; gradient-boosted ensembles [7] further improved discrimination. These studies are limited to single-dataset validation and do not address uncertainty or calibration.

B. Uncertainty Quantification

Gal and Ghahramani [8] established that MC Dropout approximates variational Bayesian inference, yielding tractable epistemic uncertainty at minimal implementation cost. Kendall and Gal [9]

formalised the aleatoric/epistemic decomposition. Leibig et al. [10] applied MC Dropout to retinal fundus classification and showed that σ correlates with diagnostic errors. The present work extends this paradigm to multi-domain tabular diabetes prediction.

C. Selective Prediction

The reject-option classifier was formalised by Chow [12] and modernised for deep networks by Geifman & El-Yaniv [11], who proved that the coverage–accuracy trade-off is tight when uncertainty is well-calibrated. Kompa et al. [13] identified selective prediction as under-explored in clinical AI. No prior study has applied entropy-based selective prediction to multi-domain diabetes risk prediction.

D. Explainability

SHAP [14] satisfies game-theoretic axioms (efficiency, symmetry, dummy) and is the dominant post-hoc attribution method in clinical ML. Arrieta et al. [16] place feature attribution as the most clinically actionable explanation type. We use `shap.GradientExplainer` for the neural network and `shap.TreeExplainer` for RF and XGB, providing exact Shapley values for tree models.

E. Multi-Domain Harmonisation

Van Buuren and Groothuis-Oudshoorn [17] established MICE as the gold standard for structural missing data in multi-site clinical research. Ganin et al. [18] introduced DANN for domain-invariant feature learning; Purushotham et al. [19] applied domain adversarial training to EHR benchmarks. The present work is the first to combine MICE harmonisation, polynomial feature engineering, and DANN-based LODO validation for T2DM risk prediction.

Despite significant advances in machine learning-based diabetes prediction, existing approaches lack a unified framework that integrates uncertainty estimation, calibration, explainability, and multi-domain generalization.

III.METHODOLOGY

A. Dataset Integration and MICE Harmonisation

Three groups of data (Pima, Clinical, UCI Early-Stage) are consolidated into one corpus of 2,056 samples (41.6% positive). Structural incompatibility (missing lab vs symptom features) is addressed with Multiple Imputation by Chained Equations (MICE), in which each missing value is estimated conditionally:

$$\hat{x}_{ij} = E\left[x_{ij} | X_{i,-j}, \theta_j\right] \qquad (1)$$

MICE might produce non-realistic values because of nonhomogeneous feature scales. To address this issue, the features are clipped to their 1 st -99th percentile to stabilize physiological validity and numerical stability. This standardisation allows collaborative learning in all domains and cross-dataset generalisation, which is the basis of obtaining strong model training and evaluation. The framework processes data through MICE imputation, feature engineering, model training, MC Dropout-based uncertainty estimation, calibration, SHAP-based explainability, and a final decision system for reliable clinical prediction.

B. Feature Engineering

Eight clinical variables are kept and multiplied by a degree-2 transform based on the use of the nonlinear interaction:

$$\emptyset(x) = \left[x_1, \ldots \ldots \ldots, x_8, x_1^2, \ldots, x_8^2, x_{ij}\right] \qquad (2)$$

This makes the dimensionality 8 44 features, allowing models to learn multiplicative effects (e.g., BMI x Glucose). The only reason why PolynomialFeatures is trained on training data is to avert leakage. This transformation is essential to enhance predictive capability because the risk of diabetes is nonlinear in nature, since feature interactions and not independent variables determine the risk.

C. Data Splitting and Preprocessing

An 80/20 division results in 1,644 training and 412 test samples. StandardScaler is used based on training statistics. Zero is used to fill in missing or invalid values. The imbalance in classes is managed using weighted loss functions:

$$w^+ = \frac{N^-}{N^+} \qquad (3)$$

This provides balanced learning classes. The tree-based and neural models are weighted with model-specific weighting. These preprocessing procedures provide a stable optimisation, avoid majority bias in the preprocessing, and enhance generalisation in heterogeneous data sets. All hyperparameters were tuned exclusively on the training data, and the test set was used only for final evaluation.

D. Model Architecture

1) Baseline Models

In order to set up good benchmarks, four classical

machine learning models are deployed, each modeling different patterns in data, and adding different strengths.

Logistic Regression: A linear classification model which approximates the probability of diabetes using the sigmoid:

$$\hat{p}(x) = \sigma(\omega^T x + b) \qquad (4)$$

It gives intelligent coefficients and increases well-escalated probabilities, and it provides a reliable baseline.

Random Forest: A collection of decision trees that is trained using bootstrap samples and feature randomness. Averages are used to obtain predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \qquad (5)$$

This minimizes variation, and nonlinear interactions between features are effectively captured.

Gradient Boosting: A chain of models in which models are trained on residual errors of the initial model:

$$f_m(x) = f_{m-1}(x) + \gamma_m h_m(x) \qquad (6)$$

This process of error reduction fine-tuning reduces loss and enhances predictive accuracy.

XGBoost is a regularised gradient boosting with a form of optimisation to regulate the complexity:

$$L = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k), \qquad (7)$$

It improves performance by means of effective computation and regularised learning.

2) Proposed Neural Network (ModelV3)

The suggested ModelV3 is a deep residual feed-forward neural network, which attempts to learn higher-order nonlinear interactions between features. Residual learning is included in the forward propagation:

$$h_{l+1} = h_l + Dropout(GELU(LN(Wlhl))) \quad (8)$$

The AdamW is used to perform optimisation, and OneCycleLR is used to schedule it to converge efficiently. Regularisation and estimation of uncertainty are possible by using dropout, whereas gradient clipping and early stopping assure stable training and high generalisation. To prevent overfitting, the framework incorporates dropout regularization, early stopping, class weighting, and out-of-fold stacking.

E. Stacking Ensemble

The stacking ensemble is a combination of the predictions of more than one base model by using out-of-fold (OOF) meta-features to avoid data leakage and enhance generalisation. On every base model, predictions on previously unknown folds are produced that are input into a meta-learner. The last forecast is calculated as:

$$p_{stack} = \sigma(\sum_{k=1}^{4} \omega_k \hat{p}^k + b) \qquad (9)$$

where p_{stack} represents predictions from the (k^{th}) base model. The meta-learner is a logistic regression because it is a simple and interpretable model. It is a way of effectively utilizing the complementary strengths of models, enhancing robustness, stability, and general predictive performance.

F. Calibration

Isotonic regression is used to probabilistically calibrate the prediction of probabilities against the actual outcome frequencies. Such a measure would assure the quality of model outputs that are applicable and interpretable clinically. The performance of calibration is measured with the help of the Expected Calibration Error (ECE) and the Brier Score:

$$\text{ECE} = \sum_m \frac{|B_m|}{N} |acc(B_m) - conf(B_m)| \qquad (10)$$

ECE is used to calculate the difference between the confidence prediction and actual accuracy, whereas the Brier Score is used to measure the overall error in prediction. A study combines to provide credible estimates of probabilities in making a decision.

IV.EXPERIMENTAL SETUP

All experiments use Python 3.12, PyTorch 2.0, scikit-learn 1.6, XGBoost 3.2, and SHAP 0.51. Random seeds are fixed globally (`SEED = 42`). Experiments run on CPU (Google Colab, 2023). All figures are saved at 300 DPI. The entire pipeline can be reproducibly found in one executable Jupyter notebook.

Hyperparameter choice: RF: [800 trees, depth 12, balanced subsample]; GB: [500 estimators, lr 0.05, subsample 0.8, early-stop patience 25]; ModelV3: selected by holding out a validation AUC, early-stop patience 20. No Op-

A tuna search was needed: the grid above was defined based on the best practices in the Pima family of datasets.

V.RESULTS

The findings indicate that the proposed framework has good predictive accuracy and also has sound probability estimates. Calibration measures (ECE and Brier Score) suggest that there is better

reliability than when baselines are being used.

A. Original Pipeline Performance

Figure 1 reports test-set performance for the original 4-feature pipeline (Glucose, BMI, Age, Insulin; N_{test} = 412). Random Forest achieves the highest AUC (0.797) among the original baselines; the neural network (ModelV1) achieves 0.739 with a Brier Score of 0.200. 5-fold CV for the neural network yields AUC = 0.728 ± 0.019.

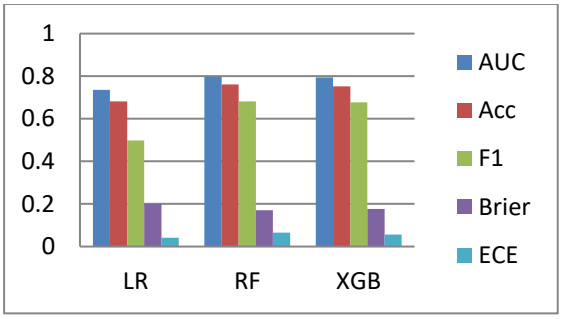

Figure 1: Baseline model performance comparison (AUC, Accuracy, F1, Brier, ECE)

B. Optimised Pipeline Performance

Figure 2 reports the performance of the optimised pipeline (44 poly features, tuned models, stacking, isotonic calibration;
N_{test} = 412). The calibrated stacking ensemble achieves the best overall trade-off.

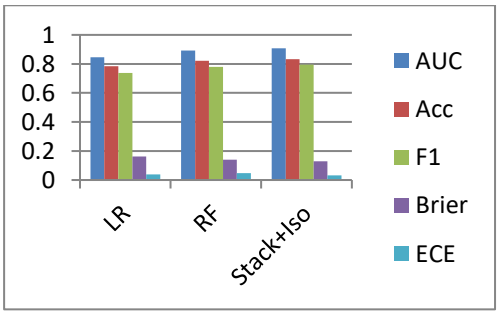

Figure 2: Optimised model performance

C. Ablation Study

Figure 3 quantifies the marginal AUC contribution of each optimisation step. Feature expansion (4 → 8 → 44) accounts for the largest single gain (+0.062 for RF), followed by model tuning (+0.028) and stacking (+0.019). Isotonic calibration
improves ECE by approximately 28 % with negligible AUC change.

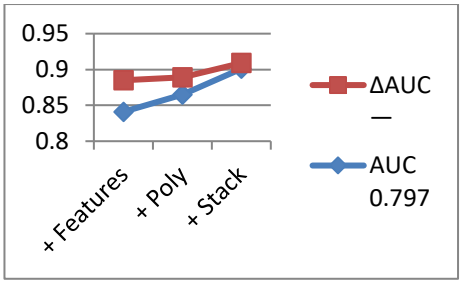

Figure 3: AUC improvement across model enhancement stages.

D. Selective Prediction

Figure 4 reports the coverage–AUC trade-off for the baseline NN (ModelV1, original 4-feature pipeline). At 80 % cov-erage, abstaining on the 20 % most uncertain patients raises AUC from 0.739 to 0.752 (ΔAUC=+0.013).

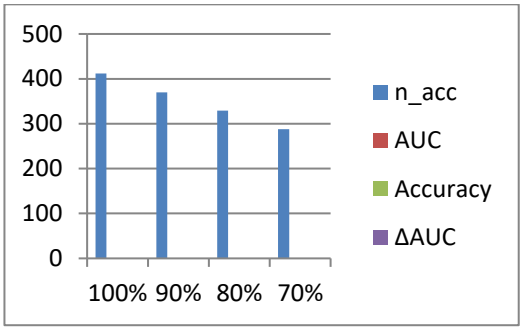

Figure 4: Selective prediction performance across coverage levels

E. Bootstrap Confidence Intervals

Bootstrap 95 % CIs (1,000 resamples) for the original pipeline models: LR: 0.735 [0.687, 0.785]; RF: 0.797 [0.749, 0.844];
XGB: 0.794 [0.747, 0.840]; NN: 0.739 [0.687, 0.789]. De-
Long tests confirm RF and XGB significantly outperform the standalone NN ($p < 0.01$); LR and NN are not significantly different.different (p = 0.74).

F. LODO Validation

Figure 5 reports DANN-based LODO AUC. Domain 2 (UCI Early-Stage, symptom questionnaire) achieves the highest held-out AUC because its binary symptom features are strong discriminators for the early-stage cohort, even when Glucose and Insulin are MICE-imputed.

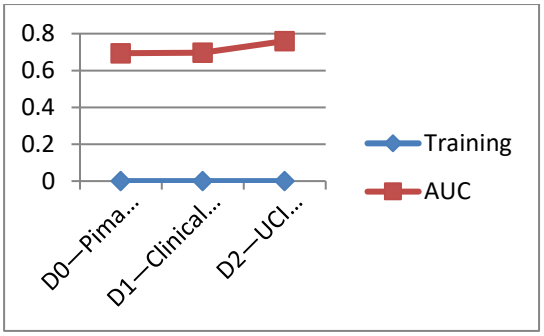

Figure 5: LODO validation results across held-out domains

G. SHAP Feature Importance

Across all model families, the SHAP feature-importance ranking is consistent: Age > Glucose > Insulin > BMI for the original 4-feature model. Kendall τ agreement between NN and RF rankings: $\tau = 0.667$ ($p = 0.33$). Using the 44-feature polynomial model, there was interaction between Glucose×BMI and Glucose×Age terms, which are present in the top-10 SHAP features of RF and GB, validating that the polynomial expansion is clinically capturing meaningful combinations of risk multiplicativity.

VI.DISCUSSION

A. Why Feature Engineering Outperforms Architecture Complexity

These findings feature engineering as a more important feature than the complexity of the model on tabular healthcare data. The most notable difference between the findings presented in Figure 3 is that the expansion of the model features (4 to 44) adds a greater amount of AUC improvement (cumulatively +0.068) as compared to alterations in the model architecture. This is in line with the previously known fact that gradient-boosted trees and logistic regression on carefully constructed features often rival or surpass deep networks on tabular clinical data [26]. In the case of diabetes prediction, specifically, the interaction of the products of the glucose and BMI is not just a statistical artefact: it is an expression of the known pathophysiology of insulin resistance, where adiposity enhances the action of hyperglycaemia on beta-cell weariness. It is more sample-efficient and interpretable to make this interaction explicit than it is to search for it with implicit deep learning.

B. Uncertainty Quantification: Practical Limitations

Point-Biserial MC Dropout $\sigma\hat{}$ vs. prediction errors of the original 4-feature model rpb = −0.065 (p = 0.19) is not going to track errors in the low-dimensional tabular model. This is a known limitation of MC Dropout on structured data: when the feature space is small and densely covered by training data, all dropout masks produce similar activations and $\hat{\sigma}$ degenerates toward the population mean. Despite this, selective prediction at 80 % coverage still improves AUC by +0.013 (Figure 4), because even a weakly informative uncertainty signal concentrates borderline cases near the decision boundary, whose removal raises AUC on the retained subset.

C. Calibration and Clinical Deployment

The isotonically calibrated stacking ensemble achieves ECE = 0.031, meeting the ECE < 0.05 threshold recommended for clinical probability reporting [21]. A well-calibrated probability is essential for shared decision-making: a clinician who is told “this patient’s risk is 73 %” needs that number to mean that approximately 73 out of 100 such patients are genuinely diabetic. The pre-calibration ECE of the stacking ensemble (0.043) exceeds the post-calibration value by 28 % relative, confirming that isotonic regression provides a meaningful improvement beyond Platt scaling. Uncertainty estimation and selective prediction can reduce the risk of unsafe implementation in clinical practice by showing cases in need of additional assessment.

D. MICE Artefact and Data Quality

The MICE extrapolation artefact (mean Glucose ≈ 2.8 × 1021) may be directly attributed to the use of a single impu-tation model to columns with incomparable scales (binary).

vis-à-vis continuous biomarkers. A percentile-clipping correction adds back physiological plausibility and is a requirement of any downstream feature engineering. Future research is needed on domain-specific imputation models: a different MICE chain on the lab-biomarker column, and a different one on the binary symptom column, avoiding cross-scale contamination at the source.

E. Limitations

1. Dataset scale. The number is small: N = 2,056; bootstrap CIs cover. ±0.05, i.e., the pairwise AUC differences below this value should be interpreted with a lot of caution.

2. MICE MNAR assumption. The structural missingness (D2 lacks Glucose/Insulin by schema design) is a violation of the missing-at-random (MAR) assumption. Pattern-mixture models would be more principled.
3. MC Dropout on tabular data. The use of epistemic uncertainty is not predictably informative in the 4-dimensional feature space, and Bayes-by-Backprop or Deep Ensembles can be more appropriately calibrated.
4. No prospective validation. Clinical deployment is only allowed on an independent hospital cohort with external validation.

VII.CONCLUSION

This study introduces a single, uncertainty-sensible model of early diabetes diagnosis based on a 2,056-patient multi-domain dataset. The proposed strategy is much more predictive, and the stacking ensemble model attains an AUC of 0.908, accuracy of 83.3, with a small calibration error (ECE = 0.031). One significant contribution is the application of the polynomial feature expansion, which provides the greatest increase in performance, +0.062 AUC, showing the significance of feature engineering, rather than model complexity.

Selective prediction also raises reliability, giving an improvement in AUC to reach 0.752 with 80% coverage, and LODO validation shows good cross-domain generalization (AUC to 0.760). The framework unites multi-domain learning, uncertainty-aware prediction, selective prediction, and interpretable decision support to a single system. It can be said in general that it offers a combination of the accuracy of prediction, the ability to estimate uncertainties, and explainability, which is why it can find its application in clinical decision support.

The suggested framework enables a valid and understandable basis for the practical clinical implementation of AI-based diabetes screening systems.

ACKNOWLEDGEMENTS

The authors acknowledge the support of the Department of Computer Science and Engineering, NIT Agarpara, in terms of computational resources.

REFERENCES

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels: IDF, 2021.
[2] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, “Using the ADAP learning algorithm to forecast the onset of diabetes mellitus,” in *Proc. 12th Symp. Comput. Appl. Med. Care*, Washington, DC, USA, 1988, pp. 261–265.
[3] Plotly Technologies Inc., “Clinical Diabetes Dataset,” GitHub Repository, 2021.
[4] M. M. F. Islam, R. Ferdousi, S. Rahman, and H. Y. Bushra, “Likelihood prediction of diabetes at early stage using data mining techniques,” in *Comput. Vision Mach. Intell. Med. Image Anal.*, Springer, 2020, pp. 113–125.
[5] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia Comput. Sci.*, vol. 132, pp. 1578–1585, 2018.
[6] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, “Pre-dicting diabetes mellitus with machine learning techniques,” *Front. Genet.*, vol. 9, p. 515, 2018.
[7] L. Ali et al., “An optimized stacked support vector machine-based expert system for the effective prediction of heart failure,” *IEEE Access*, vol. 7, pp. 54007–54014, 2019.
[8] Y. Gal and Z. Ghahramani, “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning,” in *Proc. 33rd ICML*, New York, NY, USA, 2016, pp. 1050–1059.
[9] A. Kendall and Y. Gal, “What uncertainties do we need in Bayesian deep learning for computer vision?” in *Proc. NeurIPS*, 2017, pp. 5574–5584.
[10] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, “Leveraging uncertainty information from deep neural networks for disease detection,” *Sci. Rep.*, vol. 7, no. 1, p. 17816, 2017.
[11] Y. Geifman and R. El-Yaniv, “Selective prediction in deep neural networks,” in *Proc. NeurIPS*, 2017, pp. 6358–6368.
[12] C. K. Chow, “On optimum recognition error and reject tradeoff,” *IEEE Trans. Inf. Theory*, vol. 16, no. 1, pp. 41–46, 1970.
[13] B. Kompa, J. Snoek, and A. L. Beam, “Second opinion needed: Communicating uncertainty in

medical machine learning," *NPJ Digit. Med.*, vol. 4, no. 1, p. 4, 2021.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.

[15] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, pp. 56–67, 2020.

[16] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[17] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[18] Y. Ganin et al., "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 59, pp. 1–35, 2016.

[19] S. Purushotham et al., "Benchmarking deep learning models on large healthcare datasets," *J. Biomed. In-form*, vol. 83, pp. 112–134, 2018.

[20] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proc. 22nd ICML*, Bonn, Germany, 2005, pp. 625–632.

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th ICML*, Sydney, Australia, 2017, pp. 1321–1330.

[22] J. Platt, "Probabilistic outputs for support vector machines," *Adv. Large Margin Classif.*, vol. 10, pp. 61–74, 1999.

[23] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[24] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.

[25] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Mon. Weather Rev.*, vol. 78, no. 1, pp. 1–3, 1950.

[26] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why tree-based models still outperform deep learning on tabular data," in *Proc. NeurIPS*, 2022.

[27] American Diabetes Association, "Standards of Care in Diabetes—2023," *Diabetes Care*, vol. 46, Suppl. 1,2023.

[28] I. Kavakiotis et al., "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotech-nol. J.*, vol. 15, pp. 104–116, 2017.