Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 5 | May-2026
ISSN (O) 3107-6696

Predicting Diabetes Progression Through Regression and Ensemble Learning: A Comparative Machine Learning Study

Ashhad Akhtar[1], Aquib Ali[2], Asfand Ahmad Jamalee[3]

[1]Department of Computer Science and Engineering, Integral University, Lucknow, India
[2]Department of Computer Science and Engineering, Integral University, Lucknow, India
[3]Department of Computer Science and Engineering, Integral University, Lucknow, India

Supervisor: Dr. Mohammed Akbar, Associate Professor, Integral University, Lucknow

---

Abstract - Timely and reliable forecasting of how diabetes evolves in individual patients holds substantial potential for strengthening preventive care, tailoring therapeutic decisions, and curtailing the onset of chronic complications. Binary classification schemes that merely distinguish diabetic from non-diabetic subjects are inherently limited in their clinical utility, since they collapse a multidimensional health trajectory into a single categorical label. The present investigation departs from this convention by treating disease advancement as a continuous quantitative measure, thereby generating clinically richer risk profiles. A synthetic dataset of 500 simulated patient records was constructed using physiologically plausible variables including age, body mass index (BMI), blood pressure, and serum cholesterol. The progression score was derived via a purposefully designed nonlinear equation with stochastic Gaussian perturbations. Three regression-based algorithms were trained, tested, and contrasted: Linear Regression, Random Forest Regression, and Gradient Boosting Regression. Comparative assessment relied on Mean Squared Error (MSE) and the coefficient of determination (R^2). Results consistently demonstrated that ensemble architectures — particularly Gradient Boosting Regression (MSE: 459.82, R^2: 0.9966) — outperform the linear baseline (MSE: 959.75, R^2: 0.9928), confirming the superiority of tree-based ensembles for nonlinear biomedical prediction tasks.

Key Words: diabetes progression, machine learning, regression, ensemble learning, Random Forest, Gradient Boosting, predictive healthcare.

1. INTRODUCTION

Among the chronic non-communicable diseases that impose the heaviest toll on contemporary healthcare systems, diabetes mellitus occupies a particularly prominent position. The disorder is defined by a sustained elevation of circulating blood glucose, a pathological state arising when the body's capacity to produce or respond effectively to insulin is impaired. Clinically, the disease manifests in two principal forms: Type 1, driven by an autoimmune cascade that progressively destroys pancreatic beta cells, and the considerably more prevalent Type 2, in which peripheral insulin resistance erodes glycaemic control alongside a relative decline in secretory function [1]. Both phenotypes are associated with a broad spectrum of microvascular and macrovascular sequelae including nephropathy, retinopathy, peripheral neuropathy, and cardiovascular disease [2].

Global epidemiological surveillance data confirm that diabetes-related morbidity and premature mortality continue to escalate across virtually every region of the world [3]. This worsening trajectory underscores a pressing need for computational forecasting tools capable of identifying and quantifying disease advancement well before its clinical manifestations become irreversible, thereby allowing clinicians to intervene at a stage when therapeutic impact is maximized. The emergence of electronic health records and wearable monitoring technologies has made large-scale clinical datasets increasingly accessible, catalysing the application of machine learning (ML) methods to a wide range of prognostic tasks in medicine [4].

Conventional biostatistical approaches frequently impose assumptions of linearity and feature independence that are seldom defensible in the context of complex physiological systems [5]. Machine learning, and regression-based modelling in particular, offers an alternative paradigm capable of extracting subtle nonlinear interaction patterns from multi-dimensional clinical datasets without imposing restrictive parametric constraints [6]. A critical advantage of regression frameworks over classification models is their capacity to generate granular, continuous severity scores rather than coarse binary labels, thereby furnishing practitioners with a more nuanced foundation for individualized care planning [7].

Against this backdrop, the current study systematically benchmarks three regression algorithms — Linear Regression, Random Forest Regression, and Gradient Boosting Regression — on a purpose-built synthetic clinical dataset. The central aim is to determine which algorithmic family most accurately captures continuous diabetic progression dynamics, and to establish a principled foundation for the broader deployment of ensemble methods within predictive healthcare analytics.

2. LITERATURE REVIEW

The application of machine learning to the study of diabetes has given rise to a rapidly expanding literature spanning multiple methodological traditions. Earlier contributions were largely preoccupied with binary diagnostic tasks separating individuals with confirmed diabetes from those without the

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 5 | May-2026
ISSN (O) 3107-6696

condition. Over time, however, scholars have come to recognize that quantifying disease severity along a continuous scale yields more actionable clinical intelligence than does simple categorical assignment [8]. Comprehensive reviews of the field, including the systematic survey by Kavakiotis et al. [4], have catalogued an extensive array of supervised learning techniques applied to diabetes detection and risk stratification, documenting both the methodological diversity and the performance heterogeneity that characterizes the existing literature.

Among classification-oriented studies, Sisodia and Sisodia [9] evaluated the relative efficacy of Naive Bayes, Decision Tree, and Support Vector Machine algorithms on the widely cited Pima Indians Diabetes Dataset, reporting accuracy figures that underscored the superiority of kernel-based methods for linearly separable problems. Zou et al. [10] subsequently extended such comparisons to a larger Chinese patient cohort, demonstrating that Random Forest and neural network architectures outperformed conventional logistic regression in distinguishing pre-diabetic from normoglycaemic individuals. Whilst valuable, these binary-outcome studies leave open the question of how algorithmic choices affect the prediction of continuous disease trajectories — a gap the current investigation directly addresses.

A landmark contribution to ensemble methodology was made by Breiman [11], who formalized the Random Forest algorithm. The approach aggregates predictions from a large collection of decision trees, each trained on an independently drawn bootstrap replicate of the data. At every internal node, only a random subset of input features is evaluated as candidate splitting criteria, promoting structural diversity across the forest. The resulting ensemble mitigates overfitting by averaging out the idiosyncratic errors of individual trees — an especially valuable property when working with limited or synthetically constructed datasets.

Friedman [12] introduced an alternative ensemble paradigm through the Gradient Boosting framework. Rather than parallelizing the construction of base learners, this strategy assembles them in sequential order: each successive tree is tailored to approximate the residual errors left by the existing ensemble, with the gradient of the loss function serving as the corrective signal. Through iterative refinement, the model progressively reduces systematic bias while preserving adequate predictive flexibility. Chen and Guestrin [13] subsequently proposed XGBoost — an optimized, scalable implementation of Gradient Boosting — that has come to dominate predictive modelling competitions, further demonstrating the versatility and potency of boosting-based ensemble architectures.

A growing corpus of clinical informatics research has affirmed the relative advantage of ensemble approaches over isolated algorithms when the predictive task involves continuous medical outcomes [6, 14]. Linear Regression, though widely valued for its interpretability and low computational overhead, is fundamentally constrained by its requirement of additive, linear predictor-outcome relationships. Within the highly interdependent physiological milieu of biological systems, these assumptions are routinely violated [15]. The current study extends the existing evidence base by situating all three models within an identical controlled experimental environment, making it possible to attribute performance differences to algorithmic properties rather than to incidental features of the dataset.

3. METHODOLOGY

3.1 Research Design

This investigation adopts a structured, sequential research pipeline traversing six distinct stages: synthetic data generation, definition of input features and outcome variable, partitioning of training and evaluation sets, model training, quantitative performance assessment, and cross-algorithmic comparison leading to evidence-based conclusions. At every stage, procedural choices were made with a view to preserving methodological integrity, while exclusive reliance on simulated data ensured that no ethical issues pertaining to patient privacy arose. The end-to-end workflow follows this chain: synthetic data generation → feature-target definition and train-test split → model training → evaluation via MSE and R^2 → cross-model comparison → conclusion.

3.2 Synthetic Data Generation

To sidestep the considerable ethical and regulatory barriers associated with accessing genuine electronic health records — barriers that have led several recent investigations to adopt simulation-based alternatives [16] — the study generated a synthetic patient population of 500 simulated cases. Every simulated record was characterized by four physiologically plausible predictor variables spanning the ranges described in Table 1.

Table -1: Synthetic Dataset — Feature Variables and Ranges

Variable	Range	Unit
Age	20 – 80	Years
BMI	15 – 40	kg/m²
Blood Pressure	70 – 180	mmHg
Cholesterol	100 – 300	mg/dL
Progression Score (Target)	Continuous	Synthetic nonlinear function

The quantitative target representing each patient's estimated diabetes progression score was computed through the following deterministic nonlinear function with superimposed stochastic noise:

$$Progression = 0.5 \times Age^{1.5} + 3 \times \sin(BMI/10) + 0.1 \times Cholesterol^{1.2} + \varepsilon, \quad where\ \varepsilon \sim N(0, 20)$$

The inclusion of power transformations and trigonometric functions was deliberate: these mathematical constructs embed genuine nonlinear dependencies between predictors and the outcome variable, guaranteeing that any algorithm unable to capture such complexity would incur a measurable performance penalty. This simulation strategy mirrors approaches employed in computational health studies that require controlled experimental conditions without access to real patient cohorts [17].

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 5 | May-2026
ISSN (O) 3107-6696

3.3 Data Preprocessing and Train-Test Split

Upon completing data generation, records were randomly assigned to either a training partition comprising eighty percent of the total sample (n = 400) or a held-out evaluation partition constituting the remaining twenty percent (n = 100). Because Random Forest and Gradient Boosting are invariant to the scale of input features by virtue of their tree-based decision rules, no normalization was applied prior to training these models. Standardization was, however, applied for Linear Regression to place all predictors on a common numerical footing and facilitate fair inter-model comparison. This preprocessing convention is consistent with established practice in comparative machine learning studies [14].

3.4 Evaluation Metrics

Model quality was gauged through two complementary statistical criteria. Mean Squared Error (MSE) captures the average squared discrepancy between the model's forecasts and the true progression scores, imposing a disproportionate penalty on large deviations and thereby reflecting clinical scenarios in which serious forecasting errors carry greater consequences. The coefficient of determination (R^2) expresses the fraction of total outcome variance that the model succeeds in explaining, with a value of unity denoting a theoretically perfect fit and values approaching zero indicating progressively weaker explanatory power [5]. These two metrics together provide both an absolute measure of error magnitude and a normalized index of explanatory coverage.

4. REGRESSION MODELS

4.1 Linear Regression

Linear Regression was selected as the reference benchmark for this comparative study. The model expresses the predicted outcome as a weighted summation of predictor variables according to the canonical formulation:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Parameter estimation proceeds via ordinary least squares, which minimizes the sum of squared residuals across the training set. Whilst this approach is admired for its transparency and low computational overhead, its predictive scope is intrinsically circumscribed by the linearity assumption. In the present experimental setting, where the outcome surface incorporates explicit power and trigonometric transformations, this constraint renders Linear Regression structurally ill-equipped to achieve optimal predictive accuracy. As Geman, Bienenstock, and Doursat [15] formally established, models with high structural bias inevitably underfit complex functional forms regardless of training data volume.

4.2 Random Forest Regression

Random Forest belongs to the family of ensemble learners constructed through the bagging principle originally articulated by Breiman [11]. At the training stage, a predetermined number of decision trees are each cultivated on a distinct bootstrap resample drawn with replacement from the full training set. An additional layer of randomization is introduced at every node split, wherein only a randomly chosen subset of input features is evaluated as candidate splitting criteria. This dual source of stochasticity promotes structural heterogeneity across the constituent trees. Predictions are ultimately derived by computing the mean of all individual tree outputs, a process that effectively dampens estimation variance without appreciably inflating bias. Notable operational strengths include resistance to overfitting even on small datasets, an inherent ability to model nonlinear relationships and higher-order interactions, and the production of feature importance scores that quantify each predictor's relative contribution to model performance [4, 18].

4.3 Gradient Boosting Regression

Gradient Boosting adopts a fundamentally distinct ensemble architecture based on the principle of sequential forward stagewise optimization, as formalized by Friedman [12]. The ensemble is assembled incrementally: at each iteration, a shallow regression tree is fitted not to the original target but to the negative gradient of the prevailing loss function — operationally equivalent to the current vector of prediction residuals. Each newly added tree therefore targets the specific weaknesses of its predecessors, enabling the ensemble to progressively reduce systematic bias through iterative error correction. The final predictor emerges as a weighted sum of all weak learners accumulated over all boosting rounds. Chen and Guestrin [13] subsequently demonstrated that this framework could be scaled efficiently to very large datasets through algorithmic optimizations including column subsampling, cache-aware computation, and sparsity-aware splitting. Gradient Boosting is broadly recognized as one of the most powerful general-purpose supervised learning algorithms for structured tabular data [7].

Table -2: Comparative Overview of the Three Regression Models

Model	Strengths	Limitations
Linear Regression	Interpretable; computationally efficient	Assumes linearity; poor on complex biomedical data
Random Forest	Robust to overfitting; handles nonlinearity; provides feature importance	Less interpretable; higher memory usage
Gradient Boosting	Highest accuracy; iterative error correction; captures subtle patterns	Computationally intensive; sensitive to hyperparameters

5. RESULTS

Each of the three algorithms underwent training on the 400-record training partition before evaluation against the 100-record held-out test set. The quantitative performance outcomes are consolidated in Table 3 below.

Table -3: Model Performance on the Test Dataset

Model	Mean Squared Error (MSE)	R^2 Score
Linear Regression	959.75	0.9928
Random Forest Regression	476.54	0.9964
Gradient Boosting Regression	459.82	0.9966

A discernible performance hierarchy emerged across the three models. Gradient Boosting Regression attained the most favourable evaluation profile, registering an MSE of 459.82 alongside an R^2 of 0.9966 — implying that the model accounted for approximately 99.66% of the total variance in the test set's progression scores. Random Forest Regression yielded a closely comparable outcome, with an MSE of 476.54 and an R^2 of 0.9964. Taken together, the two ensemble learners achieved roughly a fifty-percent reduction in residual prediction error relative to the linear baseline, which reported an MSE of 959.75 and an R^2 of 0.9928.

Although all three R^2 values nominally indicate strong explanatory power, the considerably inflated MSE returned by Linear Regression exposes its structural incapacity to accommodate the nonlinear functional dependencies encoded in the synthetic data. The ensemble models, by contrast, approximated these dependencies with demonstrably greater fidelity, a conclusion supported by their substantially reduced residual magnitudes. This outcome is consistent with prior literature documenting ensemble superiority in biomedical regression contexts [18, 10].

6. DISCUSSION

The experimental findings offer clear empirical corroboration of a foundational principle in supervised learning: when the data-generating process is governed by nonlinear, interacting predictor effects, ensemble tree-based models reliably surpass linear regression baselines. The present investigation instantiates this principle in the clinically meaningful domain of longitudinal diabetes management — a setting where the consequences of forecasting errors can directly influence treatment decisions and patient outcomes [2].

The performance advantage of Random Forest can be traced to its aggregation mechanism, wherein the predictions of many independently grown trees are averaged to produce a stabilized estimate. This averaging operation substantially attenuates the estimation variance that afflicts any single decision tree fitted to a finite and potentially noisy sample, a theoretical result formally characterized by Geman, Bienenstock, and Doursat [15] in their analysis of the bias-variance trade-off in statistical learning. Gradient Boosting's marginal superiority over Random Forest, though small in magnitude within the current experimental context, aligns with theoretical expectations: its sequential bias-reduction design is uniquely positioned to minimize systematic prediction errors when the data surface contains intricate curvature, albeit at the cost of heightened sensitivity to hyperparameter settings.

A necessary caveat is that all reported outcomes are predicated on synthetic rather than authentic clinical data. The exceptionally high R^2 values observed across all three algorithms are to some degree an artifact of the relatively clean and well-specified data-generating mechanism, which was free from the measurement noise, missing observations, and unobserved confounders that pervade genuine healthcare datasets [16]. It is therefore reasonable to anticipate that real-world deployment — for instance, on electronic health record repositories such as those examined by Zheng et al. [7] — would produce more modest though arguably more practically informative performance differentials.

Promising directions for subsequent research include validation against authentic patient records using systematic k-fold cross-validation, structured hyperparameter optimization through grid or Bayesian search, and the integration of post-hoc interpretability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). The incorporation of XGBoost [13] and other advanced boosting variants as additional comparative benchmarks would further enrich the evaluation landscape.

7. LIMITATIONS

A candid appraisal of this study's scope requires acknowledging several substantive constraints. The most prominent is the exclusive dependence on a computationally generated dataset rather than empirically gathered clinical records. Although the synthetic generation procedure was designed to emulate plausible physiological distributions and to embed realistic nonlinear outcome relationships, the resulting data cannot reproduce the full biological intricacy, latent confounding influences, or measurement heterogeneity characteristic of authentic patient populations [8]. The quantitative performance metrics reported herein should therefore be interpreted as indicative benchmarks rather than as projections of real-world clinical efficacy.

A second limitation pertains to the fact that every model was deployed with its default hyperparameter configuration, bypassing systematic optimization. While this choice is methodologically defensible for the purpose of a controlled comparative study, it implies that none of the algorithms was afforded the opportunity to achieve its theoretical performance ceiling. Third, the study does not incorporate cross-validation across multiple dataset partitions, preventing any formal assessment of model stability and the generalizability of the observed performance rankings. Finally, practical deployment factors — including inference throughput, probabilistic calibration, demographic equity across patient subgroups, and the computational infrastructure required for real-time clinical integration — lie outside the present study's analytical scope and warrant investigation in future work [17].

8. CONCLUSIONS

This study developed and evaluated a complete machine learning workflow for the continuous prediction of diabetes progression severity, underpinned by a purpose-built synthetic clinical dataset. A head-to-head comparison of Linear Regression, Random Forest Regression, and Gradient Boosting Regression yielded unambiguous empirical evidence that ensemble-based approaches hold a decisive predictive advantage over linear methods whenever the underlying input-

output mapping is nonlinear and multidimensional in character.

Among the evaluated algorithms, Gradient Boosting Regression demonstrated the strongest overall performance profile, achieving the lowest residual error and the highest proportion of explained variance. Its architecture — structured around sequential, residual-targeting weak learners first theorized by Friedman [12] and subsequently scaled by Chen and Guestrin [13] — appears particularly well aligned with the complexity inherent in biomedical prognosis tasks. Random Forest also delivered compelling results, offering practitioners an appealing compromise between predictive accuracy and generalization robustness [11].

The broader contribution of this investigation lies in enriching the growing evidence base that advocates for ensemble learning as a core instrument in predictive clinical analytics [4]. Framing diabetic deterioration as a continuous outcome variable, rather than reducing it to a binary category, equips healthcare providers with a more discriminating basis for patient stratification and individualized treatment planning [2]. Realizing the full translational promise of this framework will necessitate future validation on real patient data, rigorous cross-validation protocols, and the incorporation of model interpretability analyses — steps that collectively bridge the gap between computational modelling and actionable clinical application.

ACKNOWLEDGEMENT

The authors sincerely thank Dr. Mohammed Akbar, Associate Professor, Department of Computer Science and Engineering, Integral University, Lucknow, for his invaluable guidance, constructive feedback, and continuous support throughout this research work. The authors also acknowledge the Department of Computer Science and Engineering, Integral University, for providing the academic environment and computational resources necessary for carrying out this study.

REFERENCES

[1] American Diabetes Association. (2022). Standards of medical care in diabetes — 2022. Diabetes Care, 45(Suppl. 1), S1–S264. https://doi.org/10.2337/dc22-Sint

[2] International Diabetes Federation. (2023). IDF Diabetes Atlas (11th ed.). IDF. https://www.diabetesatlas.org

[3] World Health Organization. (2023). Diabetes fact sheet. https://www.who.int/news-room/fact-sheets/detail/diabetes

[4] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

[6] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning: With applications in Python (2nd ed.). Springer. https://doi.org/10.1007/978-3-031-38747-0

[7] Zheng, T., Xie, W., Xu, L., He, X., Zhang, Y., You, M., & Chen, Y. (2017). A machine learning-based framework to identify Type 2 diabetes through electronic health records. International Journal of Medical Informatics, 97, 120–127. https://doi.org/10.1016/j.ijmedinf.2016.09.014

[8] Abu-Shareha, A. A., et al. (2026). A comparative study of diabetes progression prediction techniques. Discover Artificial Intelligence, 6, 74. https://doi.org/10.1007/s44163-026-00212-1

[9] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia Computer Science, 132, 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122

[10] Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics, 9, 515. https://doi.org/10.3389/fgene.2018.00515

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

[12] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[13] 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, 785–794. https://doi.org/10.1145/2939672.2939785

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[15] Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

[16] Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2017). Accurate diabetes risk stratification using machine learning: Role of missing value and outliers. Journal of Medical Systems, 42(5), 92. https://doi.org/10.1007/s10916-018-0940-7

[17] Sarwar, M. A., Kamal, N., Hamid, W., & Shah, M. A. (2018). Prediction of diabetes using machine learning algorithms in healthcare. Proceedings of the 24th ICAC, 1–6. https://doi.org/10.23919/IConAC.2018.8748992

[18] Sohail, M. N., Jiadong, R., Uba, M. M., & Irshad, M. (2019). A multi-optimized method for diabetes classification and early warning prediction using machine learning techniques. Journal of Biomedical Informatics, 95, 103229. https://doi.org/10.1016/j.jbi.2019.103229