Dicle University
Journal of Engineering
https://dergipark.org.tr/tr/pub/dumf
duje.dicle.edu.tr

Research Article

Explainable Hybrid Deep Learning–Transformer Approach for Insulin Prediction

İlhan UYSAL[1*]

[1] Burdur Mehmet Akif Ersoy University, Bucak Zeliha Tolunay School of Applied Technology and Business Administration, Information Systems and Technologies Department, iuysal@mehmetakif.edu.tr, Orcid No: 0000-0002-6091-9110

ARTICLE INFO

Article history:

Received 23 March 2025
Received in revised form 21 July 2025
Accepted 4 September 2025
Available online 30 September 2025

Keywords:

Hybrid Modeling, Ensemble Learning, Deep Learning, Transformer Architectures, Explainable AI

Doi: 10.24012/dumf.1663768

* Corresponding author

ABSTRACT

Accurate predictive modeling is critical for enhancing patient outcomes and facilitating personalized care. This study introduces a hybrid modelling framework that combines deep learning, transformer-based architectures, and classical regression methods. The framework integrates multiple approaches, including Artificial Neural Networks, Long Short-Term Memory Networks, Convolutional Neural Networks, Random Forest, to model complex patterns in insulin biomarker data. By integrating these models into a unified framework, the approach enhances predictive accuracy while ensuring interpretability. Explainable AI techniques, including SHAP and LIME, are employed to identify key features influencing predictions, thereby promoting transparency and clinical trust. The proposed framework achieves superior performance on clinical datasets, with improved metrics such as MSE, MAE, and R^2, outperforming baseline models. Additionally, it identifies critical biomarkers associated with insulin regulation. Subgroup-level interpretations provide clinically relevant insights that inform personalized treatment strategies. This work demonstrates how advanced machine learning, coupled with explainability, establishes a robust foundation for clinical decision support systems to deliver effective and individualized patient care.

Introduction

Modern healthcare systems collect vast amounts of clinical data, which presents significant challenges for accurate analysis and interpretation of data. This challenge is more significant for chronic conditions such as diabetes and metabolic syndrome, where insulin levels significantly influence on the management of the disease. Precise estimation of insulin levels is fundamental for guiding clinical decisions and for developing tailored treatment plans. Recent evidence suggests that improvement in insulin prediction significantly enhances patient care, which highlights the need for more sophisticated insulin models and advanced modeling methods in clinical medicine [1].

Clinical data analysis has been based on traditional statistical frameworks, such as logistic regression. Although these models have a adequate performance in less complex cases, they often fail of capturing the complexity and high-dimensional clinical biomarker data. Recent comparative studies show that classical statistical methods often underperform than modern machine learning methods when dealing with advanced medical datasets. Capable of modeling non-linear relationships among numerous biomarkers, advanced deep learning frameworks, such as CNNs and LSTMs, provide effective methods for the accurate prediction of biologically complex phenomena. An illustrative example is precise prediction of insulin levels that depend on a large number of biological factors [2,3].

Insulin remains a central biomarker in metabolic studies due to its crucial role in managing diseases like diabetes. Our random forest-based feature importance analysis identified critical biomarkers associated with insulin regulation, including *AST, 25-Hydroxy Vitamin D, Glucose (Fasting), and TSH*. These findings illustrate complex interactions between liver function, thyroid health, glucose metabolism, and overall endocrine status. Addressing these diverse clinical variables together appears necessary for effectively managing complex metabolic disorders. Such observations are consistent with previous studies, which emphasize the multifaceted nature of metabolic dysfunction and support incorporating clinical and demographic data in predictive models [4, 5].

Explainable Artificial Intelligence (XAI) has emerged as a critical tool in clinical applications to enhance trust and transparency. Integrating interpretability techniques into predictive models helps clinicians better understand and accept model outputs. For example, SHapley Additive exPlanations (SHAP) is widely used to visualize how individual features contribute to insulin predictions. This

visualization aids clinicians in aligning computational results with clinical experience, fostering trust and supporting informed medical decision-making [6,7].

The subsequent stage of this research in our investigation will focus on an in-depth assessment and synthesis of predictive approaches. We focus on assessing conventional statistical methods alongside modern deep learning approaches to ascertain which one more accurately predicts insulin fluctuations. Conducting this comparative evaluation is crucial, given that previous works have documented significant variations in prediction accuracy between older and more advanced machine learning techniques [8]. Utilizing these methods of such methods to actual clinical datasets will provide empirical evidence of their value [9, 10] in guiding clinical decisions concerning the optimal analytical methods for tailored patient management.

In conducting our investigation, we have systematically applied systematic validation procedures to evaluate the reliability of the predictive frameworks. We relied on commonly accepted metrics, namely the Mean Absolute Error (MAE), Mean Squared Error (MSE), and the coefficient of determination (R-squared), to quantify the accuracy of predictions compared to actual values. Such comprehensive quantification supports the validity of the analytical process and helps in adopting predictive tools into routine clinical settings. Collectively, these contributions support the operationalization of precision medicine and provide clinicians with empirical evidence to deliver accurate, patient-centered interventions supported by computational analyses [11, 12].

In conclusion, as healthcare systems grapple with increasing data complexity, integrating deep learning and XAI provides a robust approach to enhancing patient care. By focusing on key biomarkers, such as insulin, and integrating classical and modern machine learning methods, the approach facilitates precise analysis of metabolic disorders. This integrated approach enables more accurate predictions and supports individualized therapeutic strategies. Moving forward, adopting these methods will be crucial for aligning computational advances with clinical reasoning, thereby enhancing clinical decision-making and improving patient outcomes [13,14].

The remainder of this paper is structured as follows: Section 2 describes data preprocessing and model development; Section 3 reports the results; Section 4 discusses the findings; and Section 5 concludes with clinical implications. Figure 1 illustrates the overall workflow, from raw data to interpretable predictions. The proposed hybrid ensemble model advances insulin level prediction by integrating transformer-based deep learning, classical regression, and explainable AI techniques, including SHAP [38] and LIME [41]. Unlike previous approaches, this strategy optimizes predictive performance through multi-model integration while ensuring interpretability via global and local explanation tools. It directly addresses clinical needs for actionable and trustworthy artificial intelligence (AI).

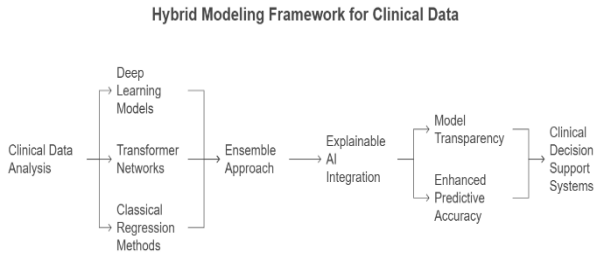

Figure 1. Model Framework

Related Works

The combined application of ensemble deep learning techniques, transformer architectures, classical regression approaches, and explainable artificial intelligence materially improves both the precision and cogency of clinical prediction models, with notable impact in diabetes management and the broader biomedical landscape [15,16]. Methods like Boosting and Random Forest steadily outperform baseline classifiers, exhibiting heightened resistance to heterogeneous and noisy clinical observations [17]. Simultaneously, convolutional neural networks and their hybrid configurations, when directed at medical imaging datasets, mitigate overfitting and foster broader generalization to external cohorts [18,19]. Transformers further expand these advantages by accommodating both structured clinical records and free-text notes, while the self-attention mechanism furnishes an intuitive basis for clinical stakeholders to interrogate model decisions [20,21]. Classical regression techniques remain important for analyzing clinical relationships and patient outcomes, especially when dealing with specific data distributions or hierarchical structures [22,23]. Additionally, explainable AI methods like SHAP and LIME play a critical role in understanding and trusting model predictions, thereby facilitating their integration into clinical practice [21]. Although transformer-based architectures have shown strong predictive performance including Badgeley et al. [21], their limited use of explainability mechanisms has restricted clinical uptake. Likewise, Afshar et al. [14] employed CNNs on clinical datasets but did not fully resolve interpretability challenges. Our approach closes these gaps by uniting hybrid deep-learning architectures with classical regression and advanced XAI frameworks, thereby improving both predictive accuracy and model transparency. In comparison to these studies, our proposed hybrid ensemble framework uniquely combines multiple deep learning architectures, classical regression models, and advanced XAI approaches. This integrated strategy aims to leverage complementary strengths of each method to improve both predictive accuracy and interpretability in insulin level prediction tasks. A concise comparative summary of important methods, domains, and key contributions—including our proposed approach—is presented in Table 1.

Table 1. Comparative Analysis of Important Studies

Reference	Method / Approach	Data / Domain	Key Contribution / Relevance
Dai et al. (2020)	Ensemble (e.g., Boosting, Random Forest)	Clinical data (various datasets)	Highlights the robustness and accuracy of ensemble methods across different healthcare datasets, addressing noise and variability through model aggregation.
Kamnitsas et al. (2018)	CNN ensemble frameworks	Medical imaging (tumor segmentation)	Demonstrates how combining multiple CNN architectures mitigates overfitting, improving generalization in medical image analysis.
Sukegawa et al. (2021)	CNN-based deep learning	Medical imaging (osteoporosis detection)	Shows CNN models' effectiveness in identifying conditions such as osteoporosis; suggests ensemble integration for improved diagnostic accuracy.
Yamamoto et al. (2020)	CNN-based approaches	Medical imaging	Explores advanced convolutional methods for enhanced diagnostic performance, reinforcing the benefits of deep learning in clinical imaging.
Badgeley et al. (2019)	Transformers + XAI	Structured and unstructured clinical data	Underscores the potential of transformer architectures for analyzing diverse clinical data, emphasizing interpretability via XAI for transparent decision-making.
Famoye & Singh (2021)	Zero-Inflated Generalized Poisson Regression	Count data in clinical outcomes	Illustrates the continued importance of classical regression methods for specialized data distributions, advocating hybrid approaches when dealing with complex clinical phenomena.
Obasohan et al. (2020)	Mixed Effects Model	Hierarchical clinical datasets	Demonstrates the use of classical statistical models to capture multi-level or hierarchical structures, providing complementary insights alongside advanced ML techniques.
Habibov et al. (2019)	Ensemble frameworks for heterogeneous data	Heterogeneous clinical data	Showcases how ensemble methods can adaptively learn from multiple data sources and handle confounding variables, offering flexibility in clinical analytics.
Proposed model	***Hybrid Ensemble (Deep Learning + Classical Regression + XAI)***	***Clinical biomarker data (single-center diabetes dataset)***	***Combines complementary modeling approaches and integrates XAI for enhanced predictive accuracy and interpretability, addressing limitations of prior single-method studies in clinical data.***

Material and Method

This study employed a clinical dataset to predict insulin levels using multiple modeling approaches. The focus is on deep learning methods, with comparisons made to classical machine learning algorithms. Hybrid ensemble models were also evaluated. This analysis is highly relevant given the growing reliance on automated techniques in medicine. Diabetes care depends heavily on accurate insulin requirement data [15,16].

Base Models with Deep Learning

ANN (MLP) Model

It is designed a model based on a Multi-Layer Perceptron (MLP) architecture, where all layers are fully connected and dropout layers are incorporated to mitigate overfitting. The model optimization step was done with the Adam algorithm, which is indicative of its popularity for regression tasks across multiple fields [25]. Recent works validate MLP's capacity to model even the most intricate datasets and thus, suitable for healthcare [15,26].

LSTM Model

The LSTM model was designed to consider the tabular data as sequential information so that the temporal insulin secretion and absorption relationships can be captured. Although the dataset contains single time-point clinical measurements without explicit temporal sequences, LSTM models were explored to capture possible latent sequential dependencies. It appears that capturing long-range dependencies and sequential patterns in glucose profiles is indeed helpful to improve prediction accuracy [27,28]. In LSTM implementations, optimization options including RMSprop are commonly chosen due to adaptive learning rates [29].

CNN Model

Sifting through the data using a one-dimensional convolutional neural network (1D CNN) enabled us to focus on the most relevant features pertaining to insulin levels. Efficient feature extraction was achieved through the stacking of Conv1D layers, flatten layers, and dense layers. SGD has also been the chosen optimizer in the recent literature [25] which supports the use of sequential data analysis [30].

Classical Machine Learning Models

Random Forest Regressor

A benchmark model and a model to evaluate feature importance was the Random Forest Regressor. Random forest models have been shown to perform particularly well when predicting outcomes related to health [15,31]. The model's robustness and efficiency in managing high-dimensional clinical data were corroborated by previous works. Random forests are usually built as an ensemble and are very useful in indicating the most important parameters that impact health metrics [30].

Hybrid Ensemble Framework

Meta-ANN (Stacking Meta-Model)

For the ensemble approach, a meta-ANN model is built using the outputs of individual deep learning models (ANN, LSTM and CNN). This stacking method utilizes various outputs from different models to improve the overall regression results. Previous research has proven the effectiveness of such hybridization to improve model accuracy and generalizability [16,32].

Hybrid (Out-of-Fold, OOF) Stacking Model

In the OOF framework, each base model is fitted on $k-1$ folds and generates predictions for the held-out fold. These out-of-fold predictions then serve as features for training the meta-model. By relying solely on unseen data for meta-level inputs, the procedure mitigates overfitting and improves generalization. This study utilizes k-fold cross-validation to obtain OOF estimates, which are subsequently used to train the meta-learner and further curb model over-learning.

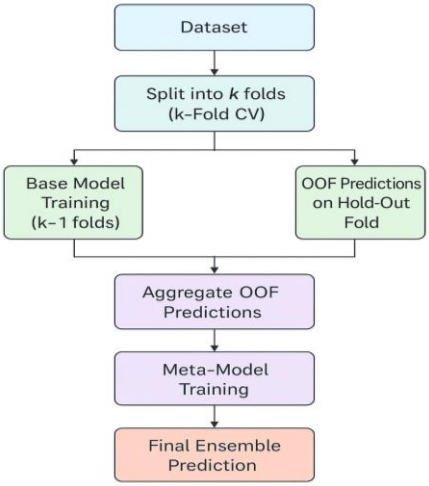

Figure 2. Diagram of Hybrid Stacking Model

Model Evaluation and Explainability

Model performance was evaluated using MSE, MAE, and R^2, each capturing distinct aspects of predictive accuracy [25]. To enhance clinical interpretability, we implemented a hybrid XAI framework combining SHAP and LIME. SHAP provides global feature importance by quantifying each variable's contribution across the dataset, while LIME offers local interpretability by approximating the model's behavior around individual predictions. Together, these methods link population-level insights with patient-specific explanations, addressing clinicians' demands for both transparency and actionable decision support [16, 24].

Application Process

The dataset from a private hospital in Antalya contains blood values of patients diagnosed with myalgia. The data set contains a total of 67 clinical variables and 2822 instances including demographic information including age, gender and biochemical parameters such as AST, ALT, glucose, TSH, ferritin. After data preprocessing steps, feature engineering was performed and the number of features was reduced to 21. The feature importance graph with random forest regressor is given in Figure 3.

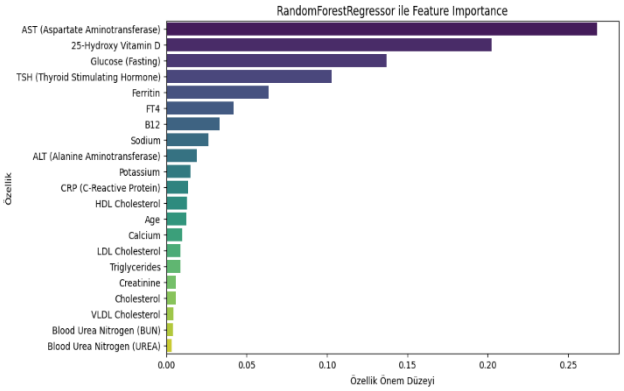

Figure 3. Feature Importance with Random Forest Regressor

During the implementation of the study, the dataset was first cleaned of missing values; columns with 90% or more missing values were removed. Remaining missing values were imputed using the Iterative Imputer method, which models missing data based on other features to provide accurate estimates. All data was scaled with Standard Scaler and then divided into a training-test set, with an 80-20 split. For deep learning models, the data was reshaped appropriately, while classical models were trained directly with scaled data. The predictions produced by the base models were used for meta-model training using stacking.

Model architectures, hyperparameters, and software libraries including their versions (Python 3.11, TensorFlow 2.12, scikit-learn 1.0.2, SHAP 0.40) are detailed in the appendix.

Results

We evaluated the predictive performance of multiple models for insulin level estimation using our clinical dataset. Table 2 compares model performances based on MSE, MAE, and R^2 score, which quantify prediction error magnitude and explained variance, respectively. The Hybrid Ensemble Model, combining MLP and Optimized Transformer architectures, demonstrated superior performance with the lowest MSE (11.43), MAE (1.43),

and highest R^2 (0.92), highlighting the advantages of integrating complementary modeling approaches.

Table 2. Comparison of Model Performances

Model	MSE	MAE	R^2 Score
Random Forest Regressor	25.3563	1.8267	0.9051
CNN+LSTM	33.3777	2.9058	0.6153
MLP	24.9893	2.6739	0.8219
Transformer	69.0674	4.5055	0.4377
Baseline Transformer	66.3250	4.4578	0.4674
Transformer + Positional Encoding	36.7644	2.5394	0.7867
Optimized Transformer	26.5052	2.3304	0.8219
Keras Tuner Optimized Model	28.6607	2.1431	0.8743

Classical machine learning methods proved remarkably effective, with the Random Forest Regressor achieving competitive results (MSE: 25.36, MAE: 1.83, R^2: 0.91) - outperforming several deep learning base models. Among individual deep learning architectures, the MLP showed stronger predictive capability (MSE: 24.99, MAE: 2.67, R^2: 0.82) compared to the CNN+LSTM model (MSE: 33.38, MAE: 2.91, R^2: 0.62), suggesting MLPs may be better suited for tabular clinical data without extensive feature engineering.

Transformer-based models exhibited varying performance levels. The baseline Transformer configurations initially underperformed ($R^2 < 0.47$), likely due to hyperparameter sensitivity and moderate dataset size. However, architectural enhancements yielded significant improvements: positional encoding increased R^2 to 0.79, while comprehensive hyperparameter optimization boosted performance further (Optimized Transformer R^2: 0.82; Keras Tuner-optimized model R^2: 0.87).

These results demonstrate that while classical algorithms like Random Forest remain robust for clinical prediction tasks, carefully designed hybrid ensembles combining optimized deep learning models achieve superior accuracy. The ensemble's performance advantage stems from its ability to capture complex, nonlinear relationships in clinical data through multiple complementary approaches.

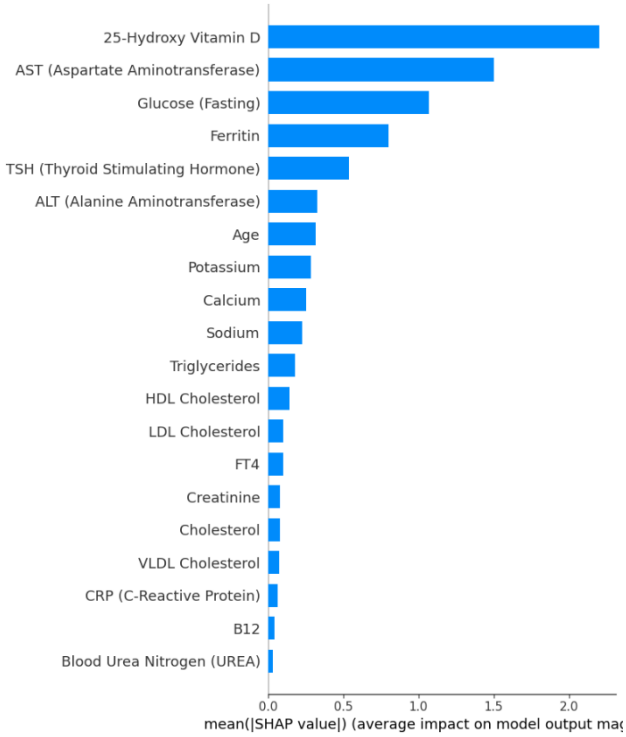

Figure 4. Average SHAP Feature Importance for Insulin Prediction Model

Using the SHAP method, the importance levels of features influencing the model's predictions are revealed. Figure 4 ranks the features based on their absolute average SHAP values, highlighting *25-Hydroxy Vitamin D* and *AST* as the most critical biomarkers for predicting insulin levels. This aligns with existing literature where vitamin D deficiency or excess significantly affects insulin regulation [35]. Other features like fasting glucose and ferritin also contribute notably, reflecting the close link between glycemic control, iron metabolism, and insulin dynamics, while *TSH* and *ALT* have smaller yet relevant impacts, with *ALT* emphasizing liver function's role in insulin metabolism.

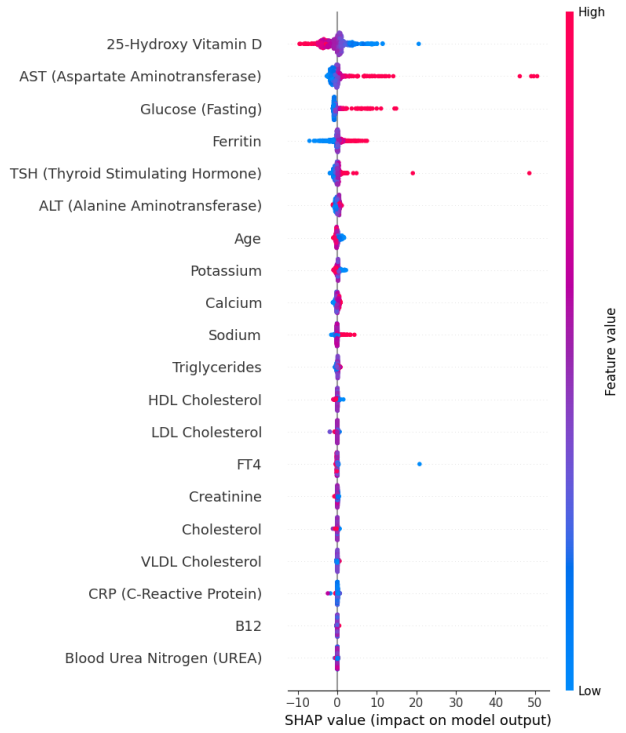

Figure 5. SHAP Bee Swarm Plot Showing Feature Value Impact on Insulin Prediction

Figure 5 complements this by illustrating how individual feature values influence insulin predictions across patients. Each point represents a patient's SHAP value for a feature, colored from blue (low feature value) to red (high feature value). For example, high fasting glucose levels (red points) correspond to positive SHAP values, increasing predicted insulin levels, whereas low glucose levels (blue points) reduce them. Similar patterns are observed for *25-Hydroxy Vitamin D* and *AST*, showing their variable contributions depending on measured values.

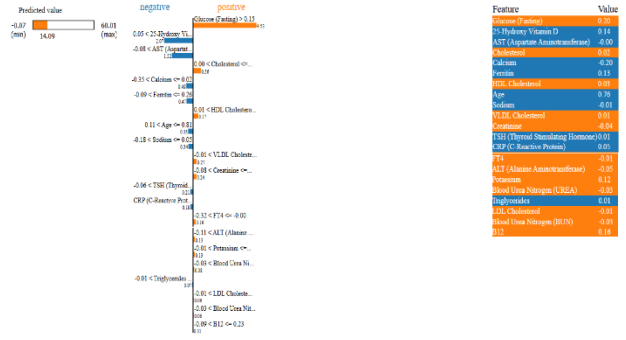

Figure 6. LIME Explanation of Feature Contributions for a Single Patient's Insulin Prediction

Figure 6, generated using the LIME method, details the positive or negative contribution of features to insulin prediction in a single example (one patient data point). This graph clearly shows that high glucose values are the strongest factor increasing insulin prediction. Some biomarkers, such as *B12* and *FT4* values, have a smaller effect on insulin levels on a sample basis and, although locally effective, have a lower effect in the general population.

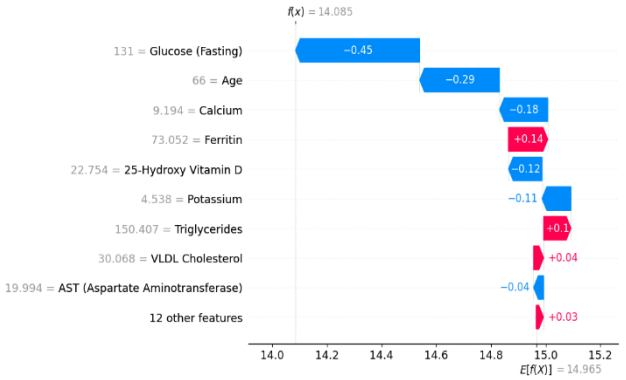

Figure 7. SHAP Waterfall Plot (Single Instance)

Figure 7 presents a SHAP waterfall plot illustrating how each feature incrementally adjusts the model's baseline prediction (14.085) to arrive at the final predicted value (14.965) for a single patient. In this instance, *Glucose (Fasting)* and *Age* exert the most pronounced negative contributions, collectively reducing the prediction by approximately 0.74 units. Conversely, *Calcium* and *Ferritin* provide moderate positive shifts, suggesting that elevated levels of these biomarkers are associated with an increased *Insulin* estimate. The contribution of *25-Hydroxy Vitamin D*—also positive—reinforces the broader observation that *25-Hydroxy Vitamin D* status plays a significant role in insulin regulation. Smaller effects, such as the negative impact of *Potassium* and the positive shifts from *Triglycerides, VLDL Cholesterol*, and *AST*, further refine the prediction. When these individual contributions are summed, the model's final prediction is slightly higher than the baseline. This granular view of how each biomarker influences the predicted *insulin* level underscores the interpretability benefits of SHAP, enabling clinicians and researchers to pinpoint the clinical factors that most strongly drive the model's decision for this particular patient.

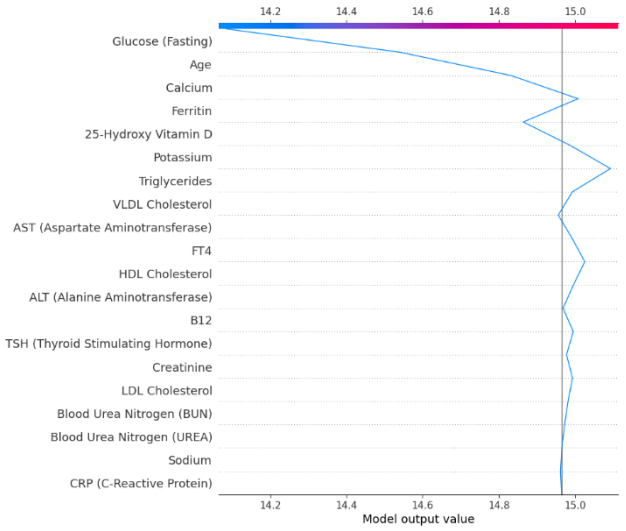

Figure 8. SHAP Decision Plot Illustrating Feature Contributions Across Individual Predictions

Figure 8 displays a SHAP decision plot that sequentially demonstrates how each clinical feature modifies the model's predicted insulin value from an initial baseline (approximately 14.2) to the final output (near 15.0) for a single instance. The horizontal axis represents the model's prediction scale, while each step in the plot shows the incremental contribution—positive or negative—of a specific biomarker:

- *Glucose (Fasting)* and *Age* appear as the top contributors, with their combined influence shaping the initial shift from the baseline.
- *Calcium, Ferritin, 25-Hydroxy Vitamin D, and Potassium* provide additional refinements, indicating that variations in these features can further raise or lower the predicted insulin level.
- The subsequent features, including *Triglycerides, VLDL Cholesterol, AST*, and others, exert smaller but still meaningful effects, cumulatively guiding the model to its final prediction.

By illustrating each feature's incremental impact, the decision plot offers a transparent view of the model's internal reasoning. Clinically, it underscores how multiple biomarkers—ranging from glucose metabolism to mineral balance—interact to influence insulin levels, thereby offering a nuanced perspective for personalized patient management.

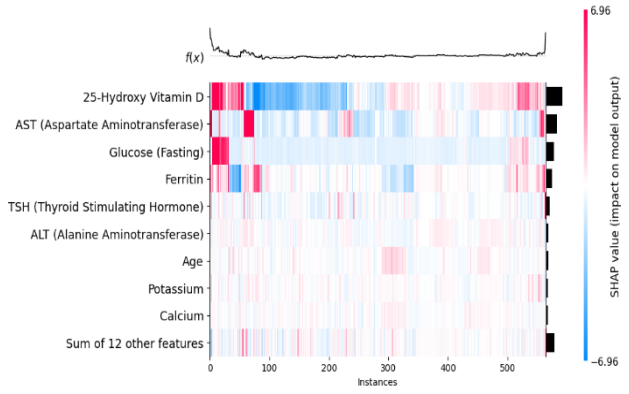

Figure 9. SHAP Heatmap Displaying Feature Impact on Model Output Across All Instances

In Figure 9, the visualization is organized into columns representing individual cases, while the rows correspond to clinical features ranked by their overall significance to the model's output. The color scale used conveys both the size and sign of each feature's influence on the resultant insulin value: saturated red denotes a feature positively influencing insulin predictions, and saturated blue denotes its negative impact. The leading feature, serum 25-Hydroxy Vitamin D, exhibits a noticeable spread of red and blue colors across the observation columns, suggesting that alterations in vitamin D status may critically impact insulin estimation according to the underlying risk profile of concurrent biomarkers. Similar spread is evident for aspartate aminotransferase, fasting glucose, and ferritin, each demonstrating distinct patterns that confirm their significant effect on the variation in predictions. Lower rows in the heatmap, containing thyroid-stimulating hormone, alanine aminotransferase, and other clinical covariates, reveal segments with less intense colors yet still instrumental in marginal adjustment of the insulin predictions. The horizontal bands of color further delineate clusters of patients whose combinatorial biomarker portraits converge to similar model predictions, suggesting the presence of latent clinical subgroups influenced by specific biomarker combinations. This heatmap illustrates both the global importance and the instance-specific role of selected biomarkers, providing clinicians and researchers with understanding of the interactions between features that collectively contribute to predictive results.

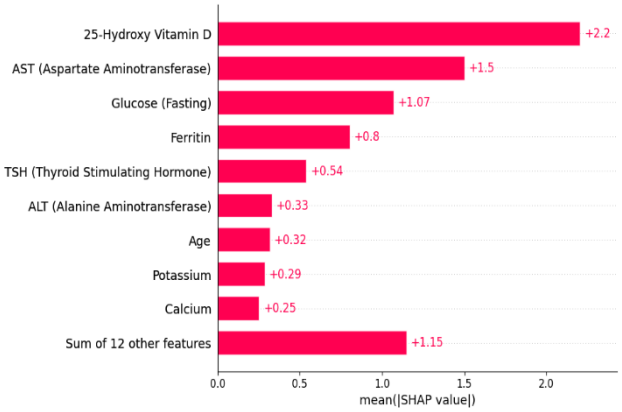

Figure 10. SHAP Bar Plot Showing Average Feature Importance in Insulin Prediction

In Figure 10, each horizontal bar indicates the mean absolute SHAP value for a given feature, illustrating its overall impact on the model's insulin predictions. The highest bar corresponds to *25-Hydroxy Vitamin D*, which stands out as the most influential predictor, with a mean SHAP value of +2.2. This suggests that variations in vitamin D levels produce larger shifts in the model's output than any other single biomarker. Following closely, *AST* and *Glucose (Fasting)* exhibit substantial mean SHAP values, highlighting their critical roles in shaping insulin estimates—an observation aligned with established clinical knowledge regarding liver enzymes and glycemic control. *Ferritin, TSH,* and *ALT* also show meaningful contributions, indicating that metabolic and endocrine factors collectively inform the model's decisions. Lower-ranked features, including *Potassium and Calcium,* have more modest mean SHAP values but nonetheless refine the prediction. The aggregated "*Sum of 12 other features*" category demonstrates that while individually less influential, a group of features still exerts a combined effect on insulin level estimation. Overall, this visualization underscores the multifactorial nature of insulin regulation, highlighting the diverse set of biomarkers—spanning vitamin levels, hepatic function, and mineral homeostasis—that the model deems most pertinent in forecasting insulin concentrations.

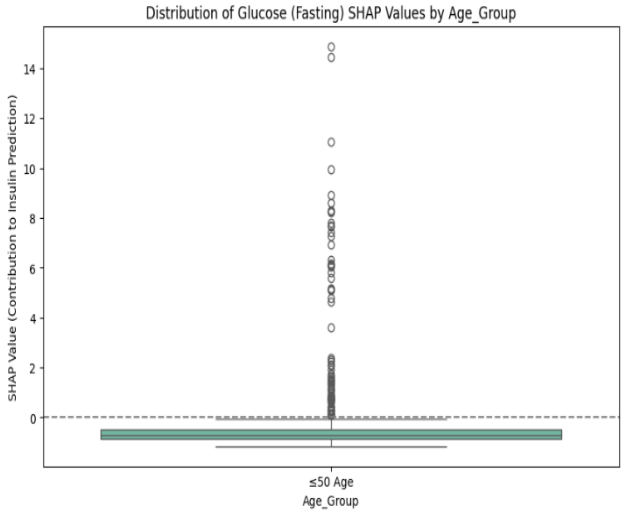

Figure 11. Distribution of Glucose (Fasting) SHAP Values by Age Group

In Figure 11, the distribution of SHAP values for fasting glucose across age groups reveals a wider spread of values in the younger subgroup (≤50 years), with several positive outliers strongly contributing to insulin prediction. This suggests that glucose levels may have a more variable influence on insulin prediction among younger individuals compared to older ones. The median SHAP value appears closer to zero, indicating that while some individuals show strong influence, the overall effect is moderate.

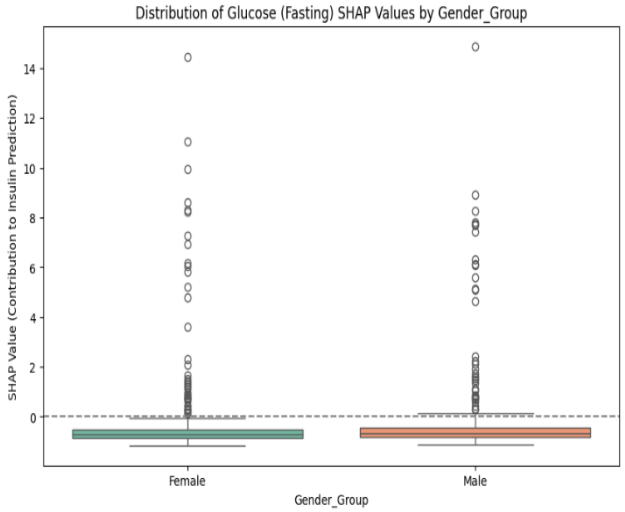

Figure 12. Distribution of Glucose (Fasting) SHAP Values by Gender Group

In Figure 12, the SHAP value distributions for fasting glucose are similar between females and males, with overlapping medians and ranges. Although outliers with high positive contributions exist in both genders, no substantial difference in the overall impact on insulin prediction is evident. This is consistent with the statistical test result ($p = 0.1842$), indicating no significant difference between genders in the predictive contribution of fasting glucose.

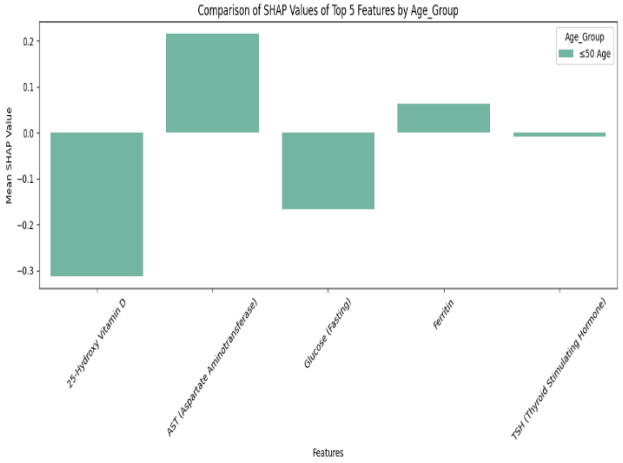

Figure 13. Comparison of SHAP Values of Top 5 Features by Age Group

In Figure 13, a comparison of the mean SHAP values of the top five features across age groups reveals notable differences. The younger group (≤50 years) exhibits higher mean SHAP values for *AST* and *25-Hydroxy Vitamin D*, suggesting these features contribute more strongly to insulin prediction in this subgroup. Conversely, the older group's influence is comparatively lower for these features. *Glucose (Fasting)* maintains a moderate effect across both groups, suggesting it is an important predictor regardless of age.

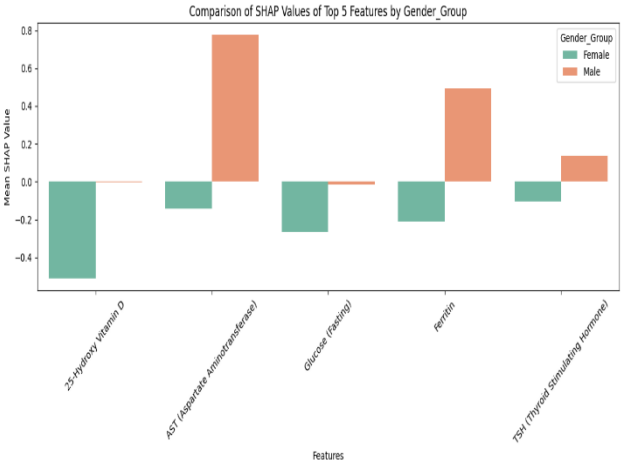

Figure 14. Comparison of SHAP Values of Top 5 Features by Gender Group

Figure 14 demonstrates distinct patterns in the comparison of mean SHAP values for the top features between genders. Males exhibit higher mean SHAP contributions for *AST* and *Ferritin*, while females show a stronger influence of *25-Hydroxy Vitamin D*. *Glucose (Fasting)* shows a relatively consistent effect in both groups. These differences imply that the model's explanation for insulin prediction varies subtly between males and females, highlighting potential biological or clinical differences in feature relevance.

Discussion

Despite the excellent performance achieved by the hybrid ensemble model, several issues remain that warrant further attention. Furthermore, it is essential to clarify whether the dataset inherently contains temporal or sequential structures that justify the use of recurrent architectures such as LSTM. If the data exhibits temporal dependencies, leveraging LSTM could provide meaningful advantages; otherwise, Transformer-based or other non-sequential models might require alternative design considerations [37].

Moreover, the success of the stacking methodology is highly contingent on the quality of out-of-fold predictions and the design of the meta-model. Transformer-based architectures initially exhibited suboptimal results, underscoring their sensitivity to hyperparameters such as learning rate, dropout, and attention mechanisms. These findings point to the necessity of advanced optimization techniques, such as Bayesian search, to unlock their full potential. Moreover, although combining MLP and optimized Transformer outputs yielded strong results (R^2 = 0.9157), further gains might be achieved by incorporating additional learners—like Random Forests or other deep models—offering diverse perspectives and improving robustness. Additionally, employing regularization strategies—such as L2 regularization or dropout at the meta-model level—could mitigate overfitting risks, leading to more robust and reliable predictions.

XAI analyses confirmed the importance of biomarkers including *25-Hydroxy Vitamin D, AST*, and *Glucose (Fasting)* in predicting insulin levels. Notably, subgroup analyses demonstrated statistically meaningful variations in feature importance across demographic strata, highlighting the need for personalized approaches tailored to demographic differences. For example, fasting glucose exhibited greater variability in predictive influence among younger patients (≤50 years), indicating heterogeneous metabolic profiles or disease manifestations in this subgroup. Gender-based comparisons, while showing no statistically significant difference in fasting glucose contribution, identified subtle but potentially clinically relevant distinctions in features including AST and Ferritin. These findings underscore the importance of integrating subgroup-specific interpretability within predictive modeling frameworks to enhance clinical transparency and support personalized medicine [38, 39]. Future work should strive to incorporate these interpretability tools more seamlessly to provide clinicians with actionable insights alongside high model performance, ultimately fostering trust in automated decision support systems [40].

The study's single-center dataset poses limitations in terms of generalizability. Broader validation across multicenter cohorts with diverse clinical profiles is essential to ensure robustness and external applicability. Future studies are needed to validate these models on larger, multicenter datasets with diverse demographic and clinical characteristics. In addition to technical improvements, ethical and practical considerations are crucial for clinical deployment. Ensuring patient data privacy, addressing potential algorithmic bias, and evaluating the clinical consequences of false predictions are essential steps for safe integration into decision support systems. These safeguards will enhance clinician trust and ensure responsible application of the model in diverse healthcare settings. Additionally, expanding the scope to include other relevant clinical outcomes, such as hypoglycemic events and long-term complications, could provide a more comprehensive understanding of disease progression and model utility in real-world settings [40].

In conclusion, our study demonstrates that hybrid ensemble methodologies, which integrate advanced deep learning architectures with classical regression techniques and robust XAI approaches, offer a promising avenue for predicting insulin levels in clinical data. Nonetheless, further optimization of model parameters, incorporation of additional base learners, explicit consideration of temporal data structure, and expansion of datasets are essential next steps to improve model generalizability and clinical utility. Additionally, embedding subgroup-specific explainability analyses will be critical for translating predictive models into trustworthy, personalized clinical decision support tools, ultimately contributing to more effective patient care [37, 38].

Conclusion

This study introduces a hybrid ensemble framework that outperforms individual models in predicting insulin levels from clinical data by combining deep learning and classical methods. However, several limitations remain. The dataset was sourced from a single center, which restricts the generalizability of the findings. Further refinement of model hyperparameters is needed, and incorporating additional biomarkers, including genetic and epigenetic factors, could enhance predictive power. Validation on larger, multicenter cohorts and the integration of real-time data streams, including continuous glucose monitoring, will be crucial to capture more complex physiological patterns. Moreover, ensuring that model explanations adapt to patient subgroups will be key to clinical acceptance, allowing for personalized and transparent decision-making. The future success of machine learning in healthcare depends on balancing high accuracy with explainable models, while addressing ethical considerations centered on patient care, including data privacy, model transparency, and potential algorithmic bias. These considerations are essential for safe and responsible use of AI in real-world clinical environments. Collectively, this study not only achieves high predictive accuracy but also provides clinically meaningful interpretations tailored to patient subgroups, representing a significant advancement in personalized AI-driven clinical decision support. Such progress holds promise for enhancing patient outcomes.

Ethics committee approval

There is no need to obtain permission from the ethics committee for the article prepared.

Conflict of Interest

There is no conflict of interest with any person / institution in the article prepared.

References

[1] Abhari, S., Kalhori, S., Ebrahimi, M., Hasannejadasl, H., and Garavand, A. (2019). *Artificial intelligence applications in type 2 diabetes mellitus care: focus on machine learning methods*. Healthcare Informatics Research, 25(4), 248. https://doi.org/10.4258/hir.2019.25.4.248.

[2] Christodoulou, E., Ma, J., Collins, G., Steyerberg, E., Verbakel, J., and Calster, B. (2019). *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models*. Journal of Clinical Epidemiology, 110, 12-22. https://doi.org/10.1016/j.jclinepi.2019.02.004.

[3] Juárez-Orozco, L., Niemi, M., Yeung, M., Benjamins, J., Maaniitty, T., Teuho, J., and Klén, R. (2023). *Hybridizing machine learning in survival analysis of cardiac pet/ct imaging*. Journal of Nuclear Cardiology, 30(6), 2750-2759. https://doi.org/10.1007/s12350-023-03359-4

[4] Yang, Z., Dehmer, M., Yli-Harja, O., and Emmert-Streib, F. (2020). *Combining deep learning with token selection for patient phenotyping from electronic health records*. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-58178-1

[5] Cheng, X., Li, S., Deng, L., Luo, W., Wang, D., Cheng, J., and Zhang, G. (2022). *Predicting elevated tsh levels in the physical examination population with a machine learning model.* Frontiers in Endocrinology, 13. https://doi.org/10.3389/fendo.2022.839829

[6] Zhang, H., Yang, Y., Yang, C., Yang, Y., He, X., Chen, C.,and Li, W. (2023). *A novel interpretable radiomics model to distinguish nodular goiter from malignant thyroid nodules*. Journal of Computer Assisted Tomography, 48(2), 334-342. https://doi.org/10.1097/rct.0000000000001544

[7] Alam, M., Islam, R., Sizan, M., and Akash, A. (2024). *The integration of machine learning in information technologies: future trends and predictions*. Journal of Computer Science and Technology Studies, 6(5), 75-84. https://doi.org/10.32996/jcsts.2024.6.5.7

[8] Song, T., Yang, F., and Dutta, J. (2021). *Noise2void: unsupervised denoising of pet images.* Physics in Medicine and Biology, 66(21), 214002. https://doi.org/10.1088/1361-6560/ac30a0

[9] Özkan, E., Orhan, K., Soydal, Ç., Kahya, Y., Tunç, S., Çelik, Ö., and Cangır, A. (2022). *Combined clinical and specific positron emission tomography/computed tomography-based radiomic features and machine-learning model in prediction of thymoma risk groups*. Nuclear Medicine Communications, 43(5), 529-539. https://doi.org/10.1097/mnm.0000000000001547

[10] Zhao, X., Yin, Y., and Bu, X. (2022). *Resilient iterative learning control for a class of discrete-time nonlinear systems under hybrid attacks*. Asian Journal of Control, 25(2), 1167-1179. https://doi.org/10.1002/asjc.2898

[11] Li, G., Ma, X., and Yang, H. (2018). *A hybrid model for monthly precipitation time series forecasting based on variational mode decomposition with extreme learning machine.* Information, 9(7), 177. https://doi.org/10.3390/info9070177

[12] Askaruly, B. and Abitova, G. (2023). *Hybrid information systems modeling technology for business process analysis based on the internet of things*. Bulletin of Shakarim University Technical Sciences, (3(11)), 19-28. https://doi.org/10.53360/2788-7995-2023-3(11)-2

[13] Wang, W. and Pai, T. (2023). *Enhancing small tabular clinical trial dataset through hybrid data augmentation: combining smote and wcgan-gp*. Data, 8(9), 135. https://doi.org/10.3390/data8090135

[14] Afshar, P., Heidarian, S., Naderkhani, F., Rafiee, M., Oikonomou, A., Plataniotis, K., and Mohammadi, A. (2021). *Hybrid deep learning model for diagnosis of covid-19 using ct scans and clinical/demographic data*., IEEE International Conference on Image Processing (ICIP), 180-184. https://doi.org/10.1109/icip42928.2021.9506661

[15] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and 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

[16] Thomsen, C., Hangaard, S., Kronborg, T., Vestergaard, P., Hejlesen, O., and Jensen, M. (2022). *Time for using machine learning for dose guidance in titration of people with type 2 diabetes? a systematic review of basal insulin dose guidance*. Journal of Diabetes Science and Technology, 18(5), 1185-1197. https://doi.org/10.1177/19322968221145964

[17] Dai, F., Meng, Y., Tan, S., Liu, P., Zhao, C., Qian, Y., and Yu, S. (2020). *Artificial intelligence applications in allergic rhinitis diagnosis: focus on ensemble learning*. Asia Pacific Allergy, 14(2), 56-62. https://doi.org/10.22541/au.159373328.85037548

[18] Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., and Glocker, B. (2018). *Ensembles of multiple models and architectures for robust brain tumour segmentation*. (pp. 450-462). Springer International Publishing. https://doi.org/10.1007/978-3-319-75238-9_38

[19] Sukegawa, S., Fujimura, A., Taguchi, A., Yamamoto, *N., Kitamura, A., Goto, R., and Furuki, Y. (2021). Identification of osteoporosis using ensemble deep learning model with panoramic radiographs and clinical covariates*. Scientific reports, 12(1), 6088. https://doi.org/10.21203/rs.3.rs-956619/v1

[20] Yamamoto, N., Sukegawa, S., Kitamura, A., Goto, R., Noda, T., Nakano, K., and Ozaki, T. (2020). *Deep learning for osteoporosis classification using hip radiographs and patient clinical covariates*. Biomolecules, 10(11), 1534. https://doi.org/10.3390/biom10111534

[21] Badgeley, M., Zech, J., Oakden-Rayner, L., Glicksberg, B., Liu, M., Gale, W., and Dudley, J. (2019). *Deep learning predicts hip fracture using confounding patient and healthcare variables*. NPJ Digital Medicine, 2(1). https://doi.org/10.1038/s41746-019-0105-1

[22] Famoye, F. and Singh, K. (2021). *Zero-inflated generalized poisson regression model with an application to domestic violence data*. Journal of Data Science, 4(1), 117-130. https://doi.org/10.6339/jds.2006.04(1).257

[23] Obasohan, P., Walters, S., Jacques, R., and Khatab, K. (2020). A scoping review of the risk factors associated with anaemia among children under five years in sub-saharan african countries. International Journal of Environmental Research and Public Health, 17(23), 8829. https://doi.org/10.3390/ijerph17238829

[24] Habibov, N., Auchynnikava, A., and Luo, R. (2019). Poverty does make us sick. Annals of Global Health, 85(1). https://doi.org/10.5334/aogh.2357

[25] Li, K., Daniels, J., Liu, C., Herrero, P., and Georgiou, P. (2020). *Convolutional recurrent neural networks for glucose prediction*. IEEE Journal of Biomedical and Health Informatics, 24(2), 603-613. https://doi.org/10.1109/jbhi.2019.2908488

[26] Fujihara, K., Matsubayashi, Y., YAMADA, M., Yamamoto, M., Iizuka, T., Miyamura, K., and Sone, H. (2021). *Machine learning approach to decision making for insulin initiation in japanese patients with type 2 diabetes (jddm 58): model development and validation study*. Jmir Medical Informatics, 9(1), e22148. https://doi.org/10.2196/22148

[27] Lee, W. (2020). *Machine learning for the diagnosis of early stage diabetes using temporal glucose profiles*. Journal of the Korean Physical Society, 78(5), 373-378 https://doi.org/10.48550/arxiv.2005.08701

[28] Muñoz-Organero, M., Queipo-Álvarez, P., and García, B. (2021*). Learning carbohydrate digestion and insulin absorption curves using blood glucose level prediction and deep learning models*. Sensors, 21(14), 4926. https://doi.org/10.3390/s21144926

[29] Tang, B., Yuan, Y., Yang, J., Qiu, L., Zhang, S., and Shi, J. (2022). *Predicting blood glucose concentration after short-acting insulin injection using discontinuous injection records*. Sensors, 22(21), 8454. https://doi.org/10.3390/s22218454

[30] Nagaraj, S., Sidorenkov, G., Boven, J., and Denig, P. (2019). *Predicting short- and long-term glycated haemoglobin response after insulin initiation in patients with type 2 diabetes mellitus using machine-learning algorithms*. Diabetes Obesity and Metabolism, 21(12), 2704-2711. https://doi.org/10.1111/dom.13860

[31] Mortazavi, B., Downing, N., Bucholz, E., Dharmarajan, K., Manhapra, A., Li, S., and Krumholz, H. (2016). *Analysis of machine learning techniques for heart failure readmissions*. Circulation Cardiovascular Quality and Outcomes, 9(6), 629-640. https://doi.org/10.1161/circoutcomes.116.003039

[32] Uyttendaele, V., Knopp, J., Stewart, K., Desaive, T., Benyó, B., Szabó-Némedi, N., and Chase, J. (2018). *A 3d insulin sensitivity prediction model enables more patient-specific prediction and model-based glycaemic control*. Biomedical Signal Processing and Control, 46, 192-200. https://doi.org/10.1016/j.bspc.2018.05.032

[33] Li, Y., Wang, H., Ye, Z., and Zhou, H. (2023). *Diabetes prediction and analysis using machine learning models*. In International Conference on Mechatronics Engineering and Artificial Intelligence (MEAI 2022) (Vol. 12596, pp. 277-283). SPIE. https://doi.org/10.1117/12.2672671

[34] Pushpavathi, K. (2024). *Diabetic drug ontology mapping for individual diabetic person and predict insulin dosage on daily basis*. Journal of Electrical Systems, 20(5s), 1801-1813. https://doi.org/10.52783/jes.2515

[35] Jung, C., Lee, M., Hwang, J., Jang, J., Leem, J., Park, J., and Lee, W. (2013). *Elevated serum ferritin level is associated with the incident type 2 diabetes in healthy korean men: a 4 year longitudinal study*. Plos One, 8(9), e75250. https://doi.org/10.1371/journal.pone.0075250

[36] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics, 8(8), 832. https://doi.org/10.3390/electronics8080832

[37] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215. https://doi.org/10.1038/s42256-019-0048-x

[38] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

[39] Tonekaboni, S., Joshi, S., McCradden, M. D., & Goldenberg, A. (2019). What clinicians want: Contextualizing explainable machine learning for clinical end use. arXiv preprint arXiv:1905.05134. https://arxiv.org/abs/1905.05134

[40] Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health, 3(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[41] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Appendix - Model Architectures and Hyperparameters Used in the Study

Model Name	Layer Structure/Architecture	Activation Functions	Optimization Algorithm	Learning Rate	Dropout Rate	Additional Parameters
MLP	3 Layers (128-64-32-1)	ReLU (Hidden Layers) Linear (Output Layer)	Adam	0.001	0.2	Epochs: 100 Batch Size: 32 EarlyStopping (Patience=10)
Transformer	CNN (Conv1D, filters=64, kernel=2) LSTM (50 units) Dense (1 unit)	ReLU (CNN Layer) Tanh (LSTM Layer) Linear (Output)	Adam	0.001	0.2	Epochs: 50 Batch Size: 32
Baseline Transformer	Embedding: Dense (32 dimensions) 2 Transformer Encoder Blocks (Head Size=16, Num Heads=4, FFN=64) GlobalAveragePooling Dense (64-1)	ReLU (Hidden Layers) Linear (Output Layer)	Adam	0.001	0.1	Epochs: 100 Batch Size: 32 EarlyStopping (Patience=10)
Transformer + Positional Encoding	Embedding: Dense (64 dimensions) 3 Transformer Encoder Blocks (Head Size=32, Num Heads=4, FFN=128) GlobalAveragePooling Dense (64-1)	ReLU (Hidden Layers) Linear (Output Layer)	Adam	0.0001	0.2	Epochs: 100 Batch Size: 16 EarlyStopping (Patience=10)
Optimized Transformer	Auto-tuned: embed_dim=[32,64,128] head_size=[16,32,64] num_heads=[2,4,8] ff_dim=[64,128,256] dropout=[0.1-0.5] learning_rate=[1e-3,1e-4,1e-5]	ReLU (Hidden Layers) Linear (Output Layer)	Adam	Tuned (Best Selected)	Tuned (Best Selected)	Epochs: 50 Batch Size: 32 RandomSearch (max_trials=10) EarlyStopping (Patience=10)
Keras Tuner Optimized Model	**Base Models:** - MLP (64-32 units) - Optimized Transformer (Embedding=64, Heads=4, FFN=128) **Meta-model:** Dense (16-8-1)	ReLU (Hidden Layers) Linear (Output Layer)	Adam	0.001	0.2 (Meta-model)	5-Fold Cross Validation Epochs: 20 (per model) Batch Size: 32