

Early Prediction of Type 2 Diabetes Using Multimodal data and Tabular Transformers

Dr. Zubair Shah

Associate Professor, College of Science and Engineering
Hamad Bin Khalifa University (HBKU), Qatar

Sulaiman Khan, Md. Rafiul Biswas (Co-authors)

Background

- Diabetes is a chronic condition that reflects body's inability to produce or respond to insulin and is characterized by abnormal glucose levels in the blood.
- WHO reports that it is the fifth most common reason for mortalities worldwide [1].
- The IDF statistics show that after every six seconds one person dies of diabetes [2].
- In 2019, approximately 425 million adults were living with diabetes; and it is projected to be 629 million by 2045 [3].

Current AI Focus

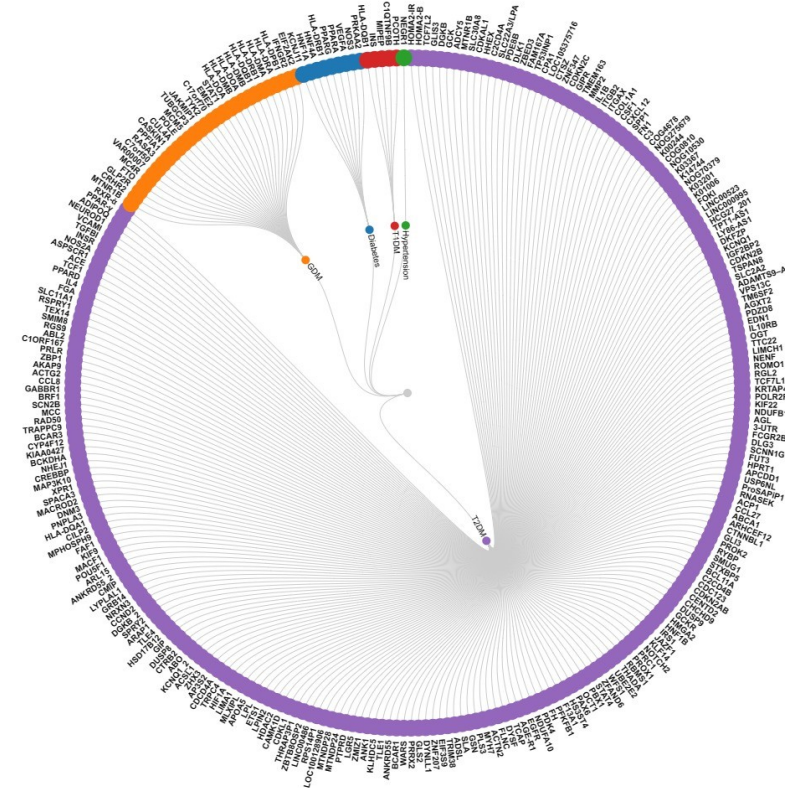
- Numerous AI/ML techniques are proposed to classify diabetic vs. non-diabetic cohorts using:
 - Electronic Health Records (EHRs)
 - Genetic data
 - Imaging data (e.g., retinal fundoscopy)
 - Other clinical/omics modalities
- Several detection models exist that integrate Bones data with other modalities to calculate risk of osteoporosis, other bones fracture or disease signals, but **none of these models used DXA data for diabetes prediction** using **longitudinal analysis**.

Clinical Reality of Diabetes Diagnosis Using AI

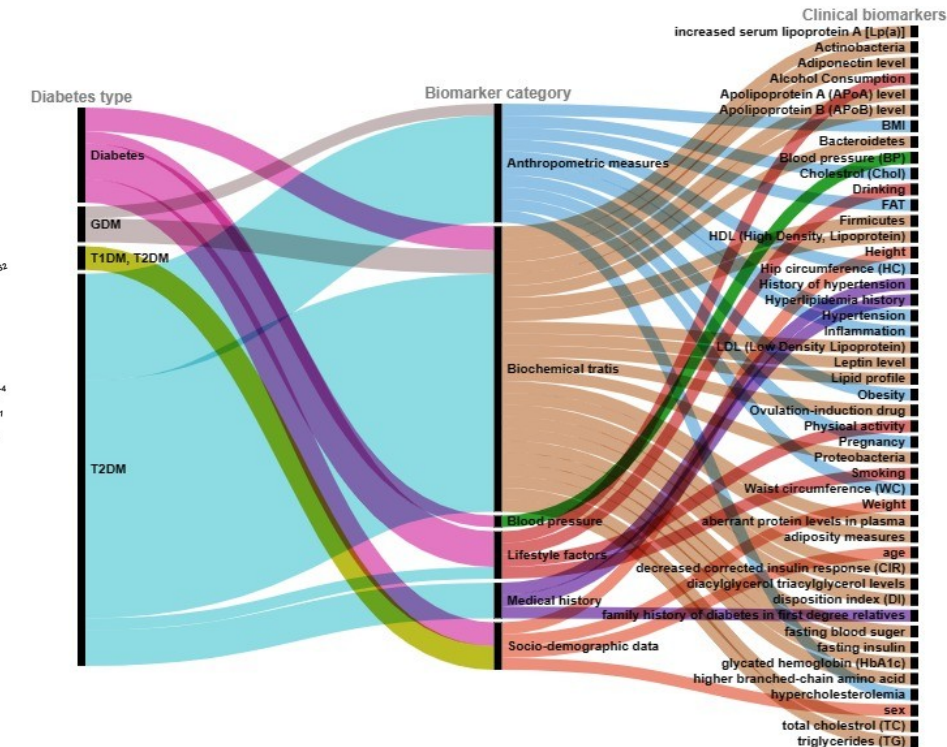
- **Diabetes Diagnosis is not the real problem.**
- Diabetes can already be detected cheaply and reliably using:
 - HbA1c
 - Fasting Plasma Glucose (FPG)
 - Oral Glucose Tolerance Test (OGTT)
- **Replacing these with AI is unlikely to happen and AI-based detection adds little clinical value in diabetes diagnosis,** since these gold-standard methods are already effective and inexpensive.
- The true challenge is **identifying who will become diabetic in the future.**

Risk Factors Associated with Diabetes

- Studies have identified diabetes risk factors, such as:
 - Genetic risk factors
 - Clinical risk factors
- However, most of these studies are **unimodal**.



Genetic factors [4]



Clinical factors [4]

Gaps in the Literature

- Many studies are **retrospective and cross-sectional**; few address prediction before onset using longitudinal data.
- **Unimodal approaches** dominate (EHR-only, genetics-only, imaging-only).
- Very limited **multimodal approaches** exist, especially those fusing **DXA data, EHR, genetic, and other clinical measures**.
- Question: Can ML models capture **longitudinal risk trajectories** of diabetes using **multimodal fusion**?

Research Aim

- Addressing this gap, we:
 - Develop a Tabular Transformer (TabTrans)-based model by fusing EHR data with bones data (DXA) to predict diabetes onset in the Qatari population
 - Perform interpretations of Models' prediction to identify longitudinal risk factors associated with Diabetes Development.

Qatar BioBank (QBB) Data

Longitudinal data

- Diabetic patients
- Pre-diabetic patients
- Normal/controlled people

Demographics

- Age
- Gender
- Nationality

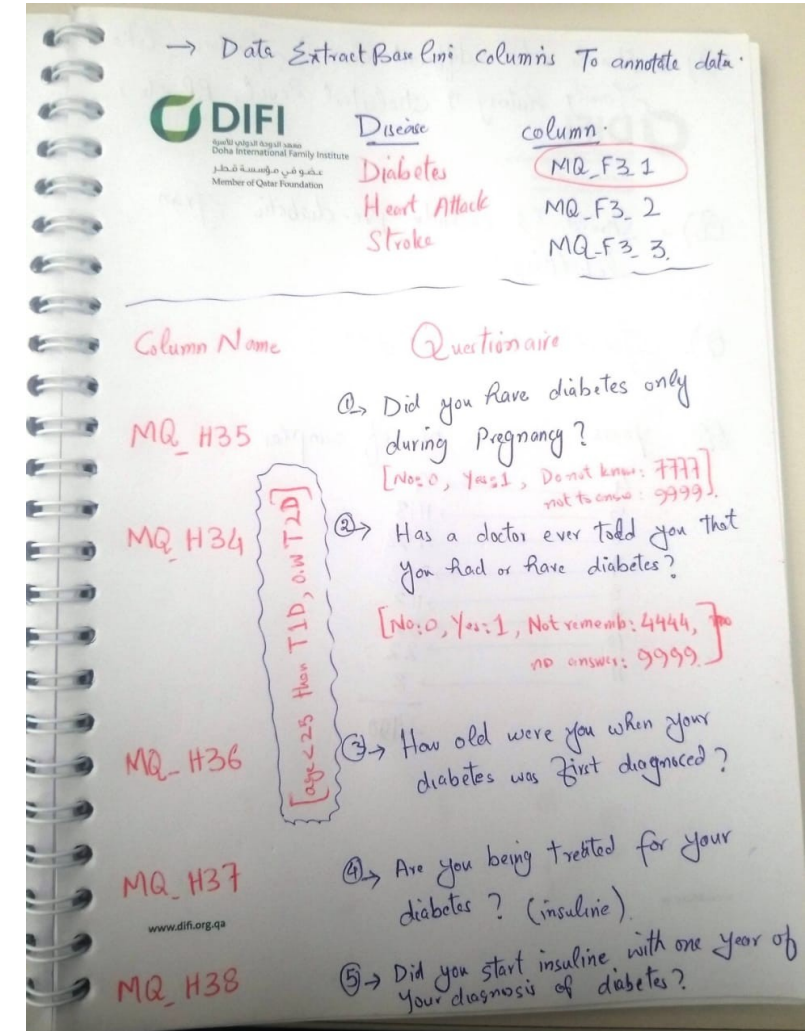
Type of data

- Retinal funduscopy images
- Clinical data

Longitudinal timeframe (years)	# of Samples
4	4
5	1118
6	798
7	293
8	113
9	50
10	22
11	2

Data Annotation

- The QBB data contains baseline and follow-up records
- We filtered the **Baseline Data** by removing pre-diabetic and diabetic participants using the following criteria:
 - (Pre-diab: IF $HbA1c \geq 5.7$ and ≤ 6.4)
 - (Diabetic: $HbA1c \geq 6.5$)
 - Self-reported pre-diabetes
 - Those who are using tablets for diabetes care (Metformine etc.)
 - start insulin within one year of your diagnosis of diabetes?
- We used only baseline data from healthy participants

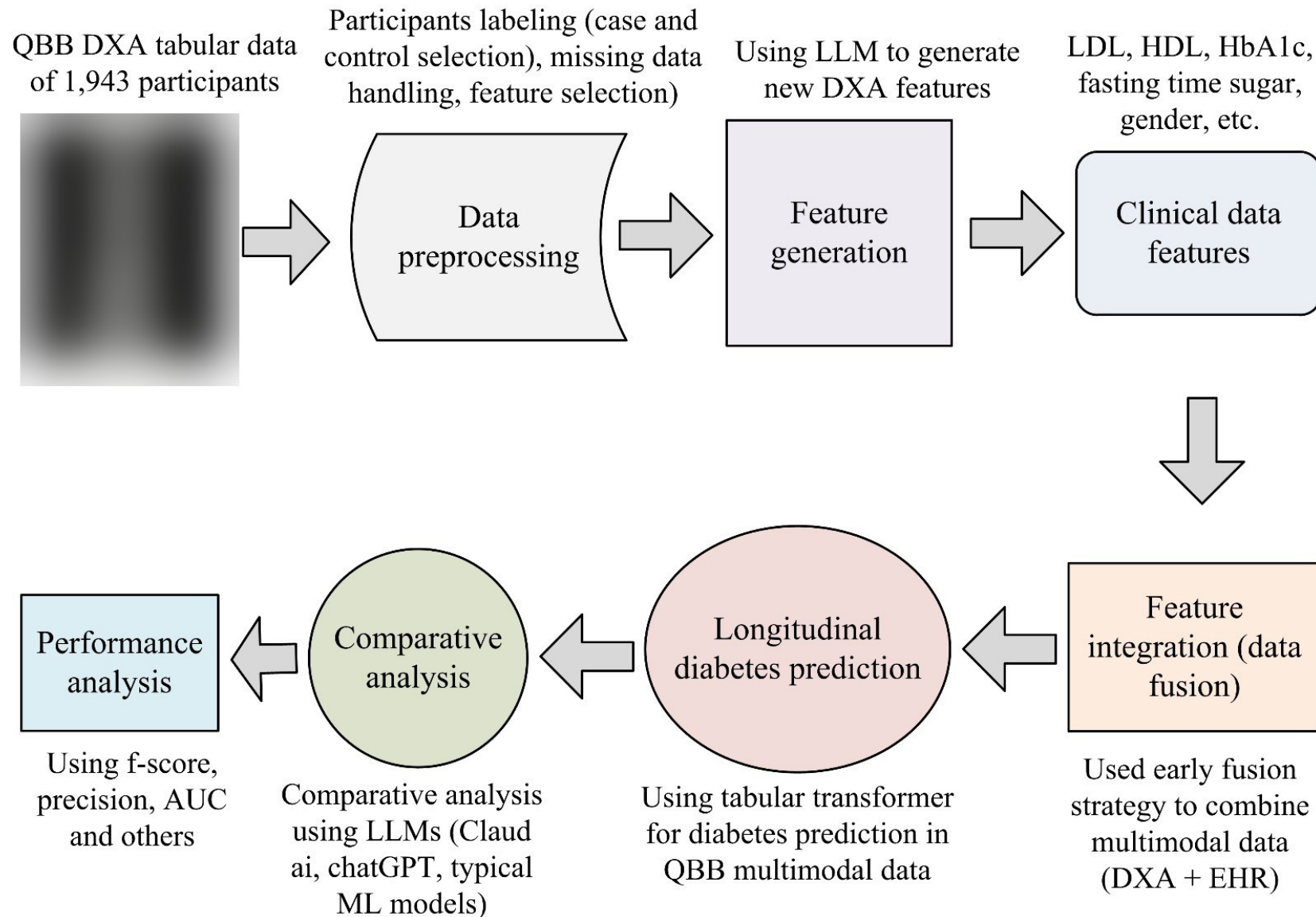


Data Annotation cont...

- A total of 1,382 participants have their longitudinal follow-up Bones Parametric data with the following specifications:
 - **Diabetic**: 279 and **Healthy**: 1,103
 - **Male**: 725 participants (diabetic: 146, healthy: 579)
 - **Female**: 657 participants (diabetic: 133, healthy: 524).
- Longitudinal Period

S. No	Longitudinal period	Number of participants
1.	4 years	4
2.	5 years	751
3.	6 years	514
4.	7 years	96
5.	8 years	17

Methodology



Prediction of Future Risk of Diabetes

● Step 1 – Labeling Strategy

- Baseline data (EHR, genetics, retinal images, etc.) is linked with **follow-up outcomes**.
- If a participant develops diabetes at follow-up → their **baseline record is labeled “future diabetic.”**
- If not → their **baseline record is labeled “non-diabetic.”**

● Step 2 – Data Preparation

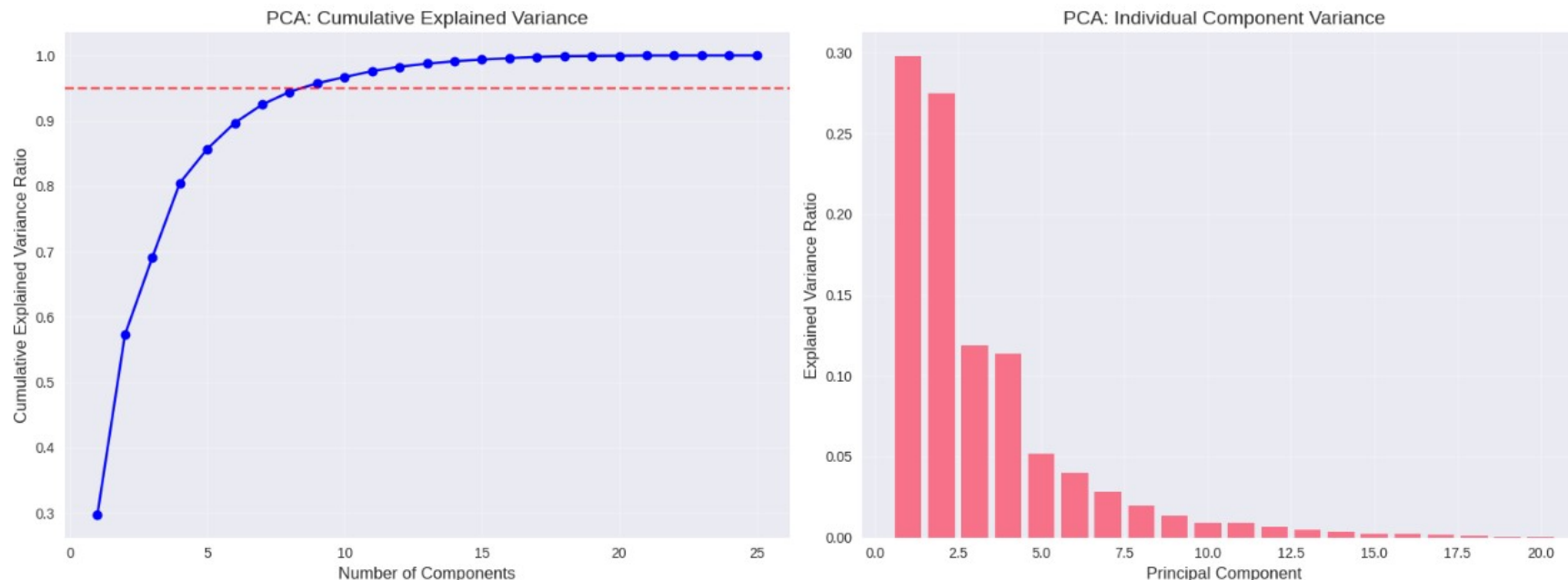
- Baseline data is now paired with future outcomes.
- Dataset is split into training and testing cohorts.

● Step 3 – Model Training & Evaluation

- AI model is trained on the **training set** to learn patterns in baseline data that predict future diabetes.
- Model performance is evaluated on the **testing set** to assess predictive accuracy.

Feature Engineering and Selection

- In DXA bones data, there are a total of **263 features** for each participants. These features contains missing values, non-correlated features.
- To explore the **structure** and **separability** of DXA-derived features, **dimensionality reduction** was conducted using **PCA** and **t-SNE**.
- PCA is applied to standardized features to **retain 95% of variance** and visualize feature contributions via biplots (**Fig**).



Continued...

- Then t-SNE is used to select only the highly correlated features using the feature map finalized using PCA.
- For datasets with fewer than 5,000 samples, t-SNE is used to visualize clustering between diabetic and non-diabetic groups, preceded by PCA reduction to 50 components when feature dimensionality exceeded a certain threshold for computational efficiency (**Tabel I**).

TABLE I
CLINICAL INTERPRETATION OF CORRELATION MAGNITUDES

Correlation Magnitude	Clinical Interpretation
$ r \geq 0.8$	Very strong clinical association
$0.6 \leq r < 0.8$	Strong clinical association
$0.4 \leq r < 0.6$	Moderate clinical association
$0.2 \leq r < 0.4$	Weak clinical association
$ r < 0.2$	Negligible clinical association

Continued...

- Finally for our finalized structural dataset, we used LLM models (to generate new features).
- Our clinical DXA diabetes prediction system implements a focused set of feature engineering methods based on established diabetes pathophysiology.
- Ten features are finalized for the final evaluation purposes.

$$\text{Central_Obesity_Ratio} = \frac{\text{Android Fat Mass}}{\text{Gynoid Fat Mass} + \epsilon} \quad (1)$$

$$\text{Visceral_Adiposity_Index} = \frac{\text{VAT Mass}}{\text{Total Fat Mass} + \epsilon} \quad (2)$$



$$\text{BMD_Coefficient_Variation} = \frac{\sigma_{\text{BMD}}}{\mu_{\text{BMD}} + \epsilon} \quad (10)$$

Data Augmentation

- Since our total bones data is: **1,382 individuals** with **T2DM cases** (n=287) and **controls** (n=1,095), which is **highly imbalanced**.
- To address this class imbalance an ensemble of advanced resampling techniques including Borderline-SMOTE, SVM-SMOTE, ADASYN, SMOTETomek, and SMOTE-ENN is evaluated.
- The optimal method was selected based on achieving a balanced class distribution while preserving data integrity, with adaptive **k-nearest neighbor** parameters automatically tuned according to minority class size.
- **Minority Class Augmentation**
 - A comprehensive augmentation pipeline is implemented to enhance minority class representation, incorporating **noise injection** ($\sigma = 0.2$), mixup augmentation ($\alpha = 0.6$), and weighted random sampling with a **10:1 minority-to-majority ratio**.
 - An augmentation factor of **5× was applied** to the minority class to ensure sufficient representation during training.

Features' Clinical Correlation Interpretation

- For **Clinical Correlations**, we used significance thresholds for correlation magnitudes using the constraints defined in Tab. I.
- **BMD Coefficient of Variation** – Bone metabolism heterogeneity is calculated using **Eq: (1)**, while linear association with diabetes is calculated using Pearson correlation as shown in **Eq: (2)**.
- A feature is selected if $|r_j| \geq 0.1$.
- **Ensemble Selection** - Features selected if they appear in any of the three methods subject to multicollinearity constraint from **Eq: 1, 2, 3**.
- The ensemble/cluster results of the feature integration is shown in Fig (next page).
- Features are also selected based on percentile thresholds values.

TABLE I
CLINICAL INTERPRETATION OF CORRELATION MAGNITUDES

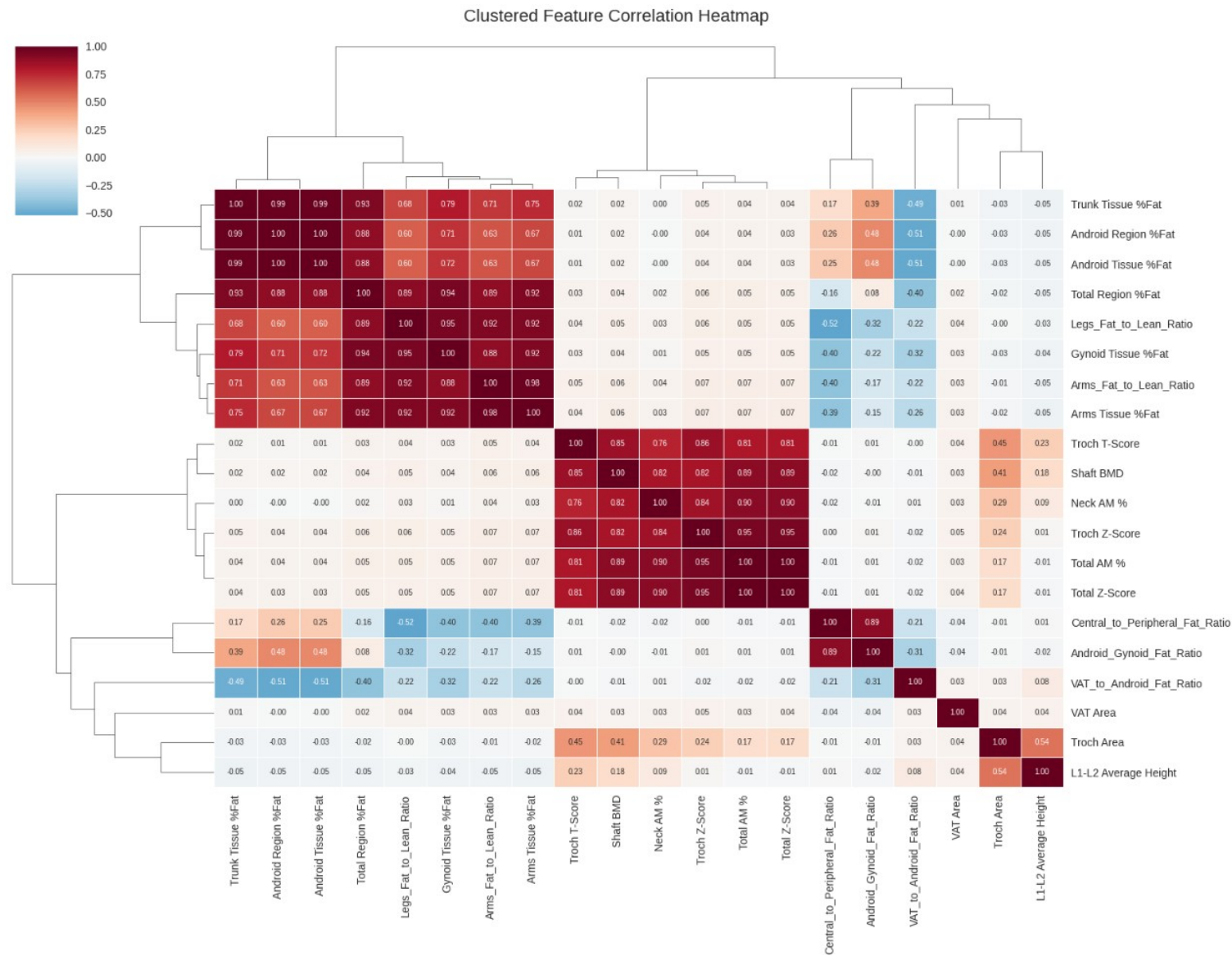
Correlation Magnitude	Clinical Interpretation
$ r \geq 0.8$	Very strong clinical association
$0.6 \leq r < 0.8$	Strong clinical association
$0.4 \leq r < 0.6$	Moderate clinical association
$0.2 \leq r < 0.4$	Weak clinical association
$ r < 0.2$	Negligible clinical association

$$\text{BMD_Coefficient_Variation} = \frac{\sigma_{\text{BMD}}}{\mu_{\text{BMD}} + \epsilon} \quad (1)$$

$$r_j = \frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (2)$$

$$\text{MI}(X_j, Y) = \sum_{x \in \mathcal{X}_j} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (3)$$

Continued...

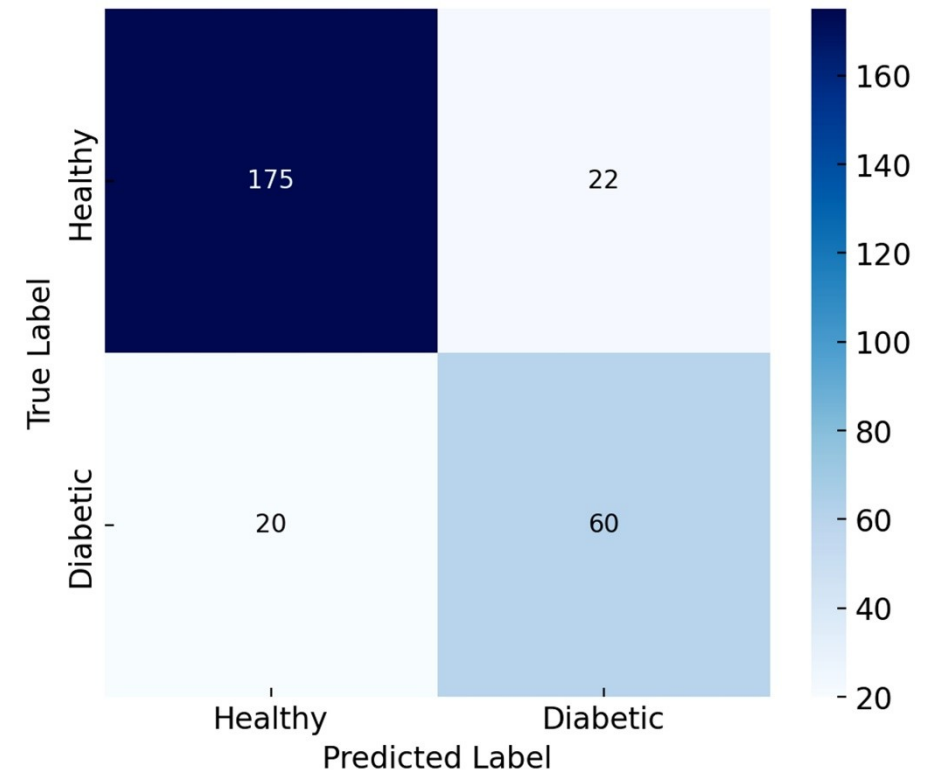


Predictions Model Evaluation

- This study employed a comparative machine learning approach to evaluate diabetes prediction performance across three distinct algorithmic paradigms:
 - **Traditional machine learning** (Eight ML models: SVM, RF, LR, AdaBoost, KNN,...)
 - **Large language models** (Claude 3.5 Sonnet, GPT-4, and Gemini Pro), and
 - **Specialized tabular transformers** (TabTrans).
- All models were trained and evaluated on identical DXA-derived feature sets to ensure fair comparison.
- Secondly, all LLM models were evaluated for diabetes prediction via standardized API interfaces to ensure uniform testing conditions.
- Each model was assigned the role of a senior endocrinologist specializing in DXA-based diabetes prediction and prompted using two strategies:
 - **Few-shot learning and**
 - **Comparative analysis.**
- Model outputs were parsed to extract structured predictions, including risk classification, probability scores (0–1), confidence levels (50–100%), and clinical reasoning.

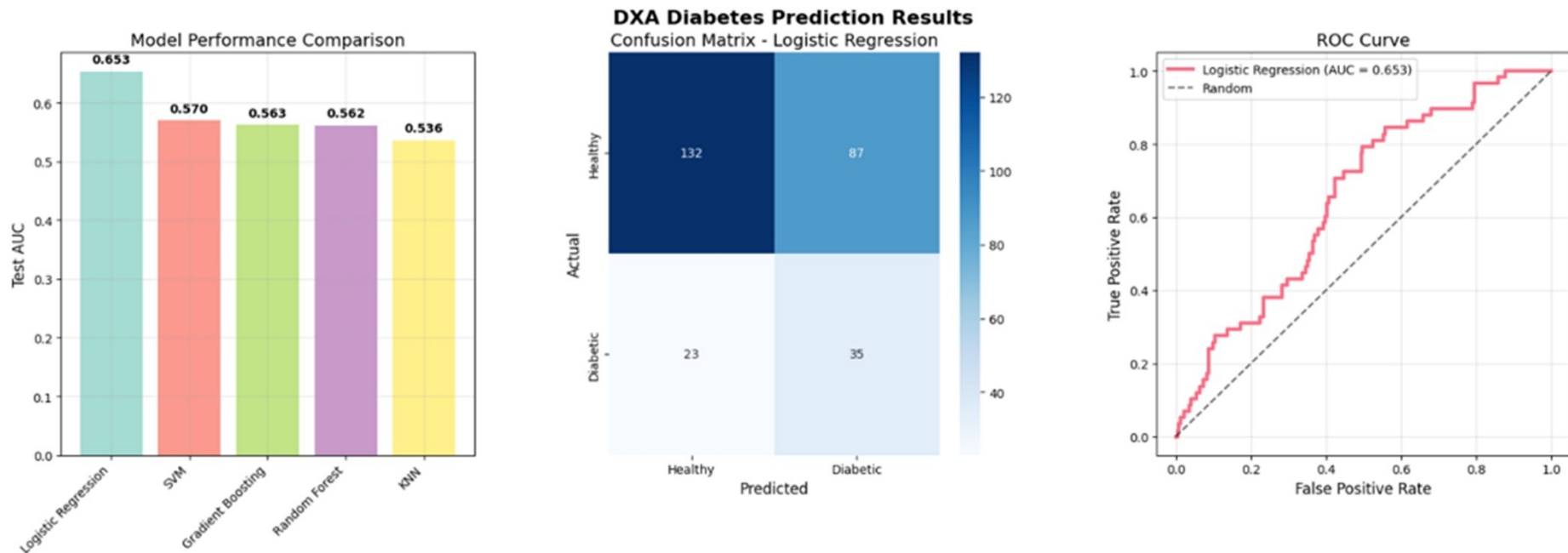
Diabetes Prediction

- For Our Model (TabTrans).
 - After applying the TabTrans model, **175 healthy** subject are correctly classified, while **22 participants** misclassified as Diabetic participants.
 - Moreover, **60 diabetic participants** are correctly classified, while **20 misclassified** as Healthy subjects as shown in Fig.
 - An overall performance of 84.8% accuracy, 73.2% precision, 75.0% recall, and 74.1% F1-score is achieved for the TabTrans model.



Diabetes Prediction

- For traditional ML models.
 - After applying traditional ML model, the highest AUC value is reported for linear regression (65.3%) much smaller than TabTrans model (79.7%).
 - Overall results for the conventional ML models are shown in Figure.



Comparative Analysis

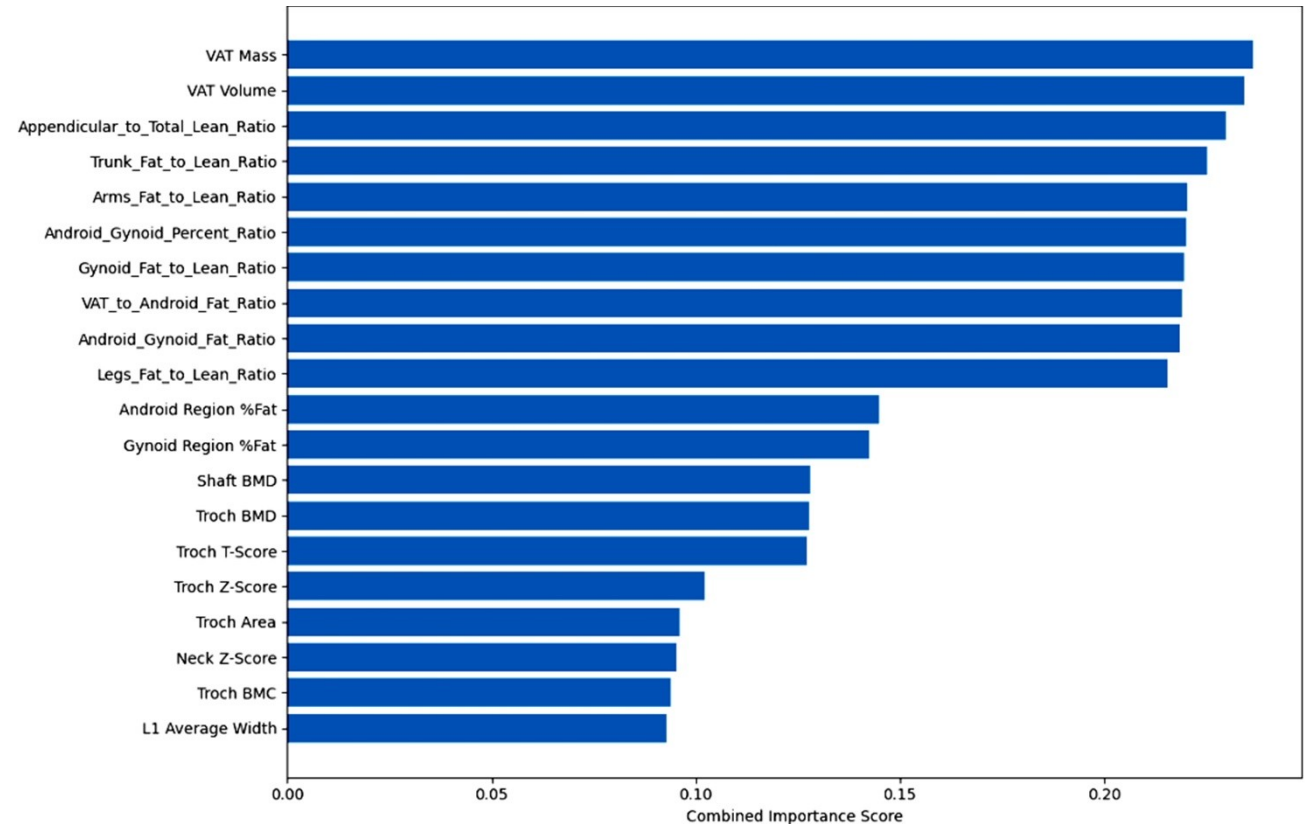
- The comparative performance of our model (TabTrans) with other benchmark models is shown in the Tabel below:

PERFORMANCE ANALYSIS OF TABTRAN VS GENERATIVE MODELS AND CONVENTIONAL ML MODELS.

Model Name	Accuracy (%)	Precision (%)	Recall (%)	ROC AUC (%)	Hyper-parameter Tuning (AUC %)
Claude 3.5 Sonnet	70.3	66.5	67.8	74.7	—
GPT-4	56.7	51.8	53.1	60.6	—
Gemini Pro	61.2	58.4	56.7	71.3	—
Logistic Regression	57.8	25.6	53.4	61.7	65.3
Support Vector Machine	54.3	48.2	54.7	56.2	57.0
K-Nearest Neighbor	46.2	41.8	47.1	46.4	53.6
Decision Tree	51.5	48.3	50.2	44.3	56.6
Gaussian Naive Bayes	48.6	41.3	47.2	45.7	52.3
AdaBoost	51.3	47.5	49.3	48.7	56.3
Random Forest	49.2	46.2	47.9	49.8	56.2
Tabular Transformers	84.8	73.2	75	79.7	59.1

Features Interpretation

- To perform feature interpretation, we used **Probabilistic Model**.
- The results are shown in Fig. **Y-axis** represents features' name, while **X-axis** represents how much each feature contributes to the model's predictions.
- The models classified the feature importance into three sections including **high contributors**, **moderate contributors**, and **least contributors**.



Limitation

- The dataset is sourced exclusively from QBB, containing only Qatari participants, which limits the generalizability to other populations.
- The sample size is relatively small (1,382 participants) and highly imbalanced, increasing the risk of model overfitting.
- The follow-up period is short, with only 17 participants having eight years of medical history, limiting insights into long-term diabetes risk factors.
- The exclusion of prediabetic individuals, and type 1 diabetes patients may affect the study's comprehensiveness.

Conclusion

- This research work presents an approach for diabetes onset prediction using artificial intelligence and DXA bones data.
- After fusing Bones tabular data features with EHR (LDL, HDL, HbA1c, demographics), we employed both conventional, LLMs, and transformer-based model to perform predictions.
- During experimental analysis, it was observed that transformer-based model (**TabTrans**) outperformed both conventional ML models and LLMs by achieving 84.8% accuracy, 79.7% AUROC, and 75% recall values.
- Through probabilistic interpretation methods, we identified key risk indicators (Vat mass, Vat volume, Ward BMD, Ward BMC, T-scores and Z-scores, Trunk and Arm Lean ratio, L1 - L4 scores) associated with diabetes development in Qatari adults.

Acknowledgements

- We thank the Qatar Biobank laboratory and spectrometry teams for providing clinical and DXA bones data and acknowledge Qatar Biobank (Project QF-QBB-RES-ACC-00181) for data access.
- We are grateful to all biobank participants for their contributions to scientific advancement.

Thank you!

zshah@hbku.edu.qa

References

- [1]. P. Balakumar, K. Maung-U, and G. Jagadeesh, "Prevalence and prevention of cardiovascular disease and diabetes mellitus," Pharmacol. Res., vol. 113, pp. 600 - 609, Nov. 2016.
- [2]. S. Pugazhenth, L. Qin, and P. H. Reddy, "Common neurodegenerative pathways in obesity, diabetes, and Alzheimer's disease," Biochim. Biophys. Acta-Mol. Basis Disease, vol. 1863, no. 5, pp. 1037-1045, 2017.
- [3]. N. Cho, J. Shaw, S. Karuranga, et al., "IDF diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045," Diabetes Res. Clin. Pract., vol. 138, pp. 271281, Apr. 2018.
- [4]. Khan, S., Mohsen, F., & Shah, Z. (2024). Genetic biomarkers and machine learning techniques for predicting diabetes: systematic review. Artificial Intelligence Review, 58(2), 41.
- [5]. Mohsen, F., Shah, U., Islam, A., Shah, Z., Belhaouari, S.B.: Multimodal deep learning for diabetic retinopathy grading: Integrating linear-radon sinograms and retinal fundus images. In: 2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP). pp. 1-5. IEEE (2024)
- [6]. Saxena, G., Verma, D.K., Paraye, A., Rajan, A., Rawat, A.: Improved and robust deep learning agent for preliminary detection of diabetic retinopathy using public datasets. Intelligence-Based Medicine 3, 100022 (2020)