JoCoSiR
Journal of Computer Science Research
ISSN: 2986-2337(Online)

‘

Classification of Health Indicators for Diabetes Mellitus Prediction Using a TabTransformer Model on Clinical Tabular Data

Al Khaidar [a], *Sri Kurnia* [b]

[a b] *Master of Information Technology Study Program, Malikussaleh University, Jl. Batam, Bukit Indah Campus - Lhokseumawe, Aceh.*

email: [a] *alkhaidarkutablang@gmail.com,* [b] *kurniazampisraf@gmail.com*

ARTICLE INFO

Keywords:
Diabetes Mellitus
TabTransformer
Deep Learning
Clinical Tabular Data
Machine Learning
Classification

IEEE style in citing this article: [citation Heading]
F. Fulan and F. Fulana, "Article Title," *JoCoSiR: Jurnal Ilmiah Teknologi Sistem Informasi*, vol. 7, no. 1, pp. 1-10, 2021. [Fill citation heading]

ABSTRACT

Diabetes mellitus is a non-communicable disease with a continuously increasing global prevalence and impacts quality of life and long-term economic burden; therefore, data-driven early detection is crucial to prevent clinical complications. This study aims to develop a diabetes prediction model using the TabTransformer architecture by utilizing a clinical dataset from Kaggle containing 100,000 patient profiles with more than 35 relevant numerical and categorical attributes. The research stages include preprocessing to remove potential leakage features, target and feature separation, numerical normalization, and categorical feature embedding. The TabTransformer model is applied for binary classification (diagnosed_diabetes) by utilizing a self-attention mechanism to capture latent interactions between tabular features, and is evaluated using accuracy, precision, recall, F1-score, and ROC AUC metrics. The results show competitive performance with an accuracy of 82.55%, a diabetes class F1-score of 0.8527, and a ROC AUC value of 0.9009, indicating the model's discriminatory ability to reliably distinguish diabetic and non-diabetic patients. Based on these results, the TabTransformer architecture has been proven effective for processing large-scale clinical tabular data and is worthy of consideration in the implementation of an artificial intelligence-based medical decision support system for predicting chronic diseases, especially diabetes mellitus.

1. Introduction

Diabetes mellitus is a non-communicable disease with a steadily increasing prevalence globally [1],[2],[3],[4]. The World Health Organization (WHO) reports that diabetes causes millions of deaths annually and contributes significantly to the health and economic burden [5],[6],[7]. This disease not only impacts patients' quality of life but also increases long-term medical costs, given its chronic nature and potential for serious complications such as heart disease, kidney failure, amputations, and vision impairment.

Early prediction of diabetes mellitus risk is becoming increasingly important to minimize complications and facilitate preventive interventions [8],[9],[10]. However, conventional approaches to identifying at-risk patients are often based on manual examinations or subjective clinical judgment, potentially leading to human bias and delayed diagnosis. The advent of modern computing technology allows for systematic and objective clinical data processing, opening up opportunities for the use of artificial intelligence in predictive health systems [11],[12],[13].

In recent years, machine learning has been used to classify diabetes-related health indicators. Algorithms such as Random Forest, Support Vector Machine, and Logistic Regression have demonstrated good performance on tabular clinical data [14],[15],[16]. However, one of the main challenges in health data is the presence of categorical features with complex interactions between variables that are often not optimally captured by traditional models.

To address this challenge, a Transformer architecture-based model, initially popular for text and sequential data, has now been adapted for tabular data through TabTransformer. This model is capable of representing embeddings in categorical features and capturing interactions between features using a self-attention mechanism [17],[18],[19], thus expected to provide more accurate and stable classification results on clinical data than traditional models [20],[21].

This study used clinical data on Diabetes Mellitus sourced from the Kaggle platform. The dataset contains health indicators such as age, blood pressure, glucose levels, body mass index, insulin, and several other clinical parameters commonly analyzed in health research. The labeling of the dataset allows for training a supervised classification model to predict a patient's diabetes status based on the indicators [22],[23].

The primary method used in this study is TabTransformer as a classification model for tabular data. This model is considered suitable because it utilizes deeper representation learning for categorical and numeric features simultaneously [24],[25]. By adopting an attention mechanism, the model is expected to be able to understand implicit relationships between clinical variables that are often difficult to detect with conventional statistical approaches. Based on the above description, this research was conducted to develop a health indicator classification system for predicting diabetes mellitus using the TabTransformer model [26],[27]. The use of this method is expected to not only improve prediction accuracy but also contribute to the development of data-driven medical decision support systems. Furthermore, the research results are expected to serve as a reference for implementing artificial intelligence in the health domain, particularly in the early prevention of chronic diseases.

2. State of the Art

2.1. TabTransformer Architecture

TabTransformer is a deep learning architecture specifically designed for processing heterogeneous tabular data consisting of a combination of numerical and categorical features [28]. This model overcomes the limitations of traditional machine learning methods like XGBoost and Random Forest, which rely on manual encoding of categorical data, by leveraging embedding and self-attention to learn contextually representative representations of categorical features. Through the embedding process, each category is converted into a fixed-dimensional numeric vector that can be learned during training, preventing the loss of information between categories, as is the case with one-hot encoding.

The main component of TabTransformer is the Transformer Block, which uses a multi-head self-attention mechanism to capture latent relationships between categorical features before combining them with numerical features. The combined representation is then passed to a Multi-Layer Perceptron (MLP) as the final classifier. This approach allows the model to adaptively extract non-linear patterns and interactions between features, resulting in more consistent performance on large-scale tabular data compared to classical models or conventional deep learning models like pure MLP without attention.

2.2. Deep Learning on Health Data Tables

The application of deep learning to tabular health data has grown rapidly due to the availability of large-scale medical datasets and the increasing complexity of clinical patterns that are difficult to model with traditional methods [29]. Various deep learning models, such as Entity Embedding MLP, TabNet, and TabTransformer, have been used for disease diagnosis, risk prediction, and disease severity classification based on clinical, demographic, and lifestyle features. The advantage of deep learning approaches lies in their ability to learn latent representations without the need for intensive manual feature engineering.

In health data, deep learning not only improves prediction accuracy but also provides generalizability to heterogeneous and noisy data. Models like TabTransformer are able to capture non-linear interactions between clinical attributes for example, the indirect relationship between age, family history, smoking patterns, and metabolic biomarkers with diabetes status [30]. This makes deep learning a relevant approach to support early detection systems, data-driven clinical decision-making, and the implementation of more precise predictive health systems.

3. Method

3.1. Diabetes Dataset

In this research, we utilized a comprehensive diabetes dataset sourced from Kaggle as the primary dataset for model development and evaluation. The dataset comprises 100,000 unique patient profiles containing more than 35 medical and demographic attributes relevant to diabetes diagnosis and risk prediction. This dataset was selected due to its clean and well-structured format, making it suitable for direct implementation in machine learning pipelines.

Each record in the dataset includes a combination of numerical attributes such as age, bmi, glucose_fasting, cholesterol, and hba1c, as well as categorical attributes including gender, smoking_status, family_history_diabetes, and employment_status. These attributes collectively represent both clinical measurements and lifestyle-related factors that are commonly associated with the development and progression of diabetes.

In addition to the feature set, the dataset provides two target variables that support supervised learning tasks, namely:

1. diagnosed_diabetes : used for binary classification (diabetic vs. non-diabetic)
2. diabetes_stage :used for multi-class classification of disease severity.

This dataset serves as a strong foundation for developing and validating machine learning models, particularly in this study where the TabTransformer architecture is applied to handle heterogeneous tabular data. A detailed summary of the attributes used is presented in Figure 1.

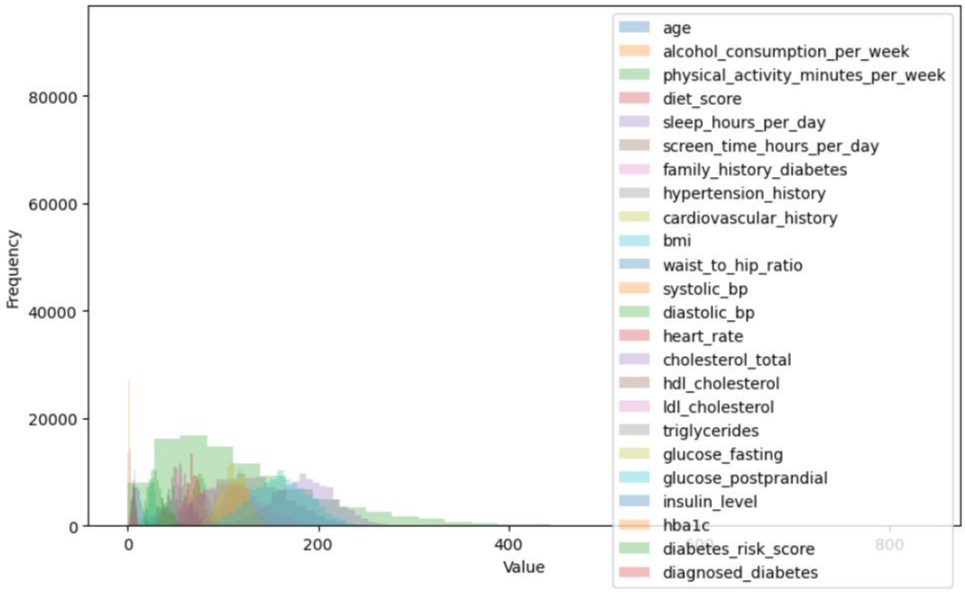


Figure 1. Dataset

3.2. Architecture Of The Tab Transformer Model

The TabTransformer model is a deep learning architecture specifically designed to handle heterogeneous tabular data with categorical and numeric features. Unlike traditional machine learning algorithms, TabTransformer leverages a self-attention mechanism to capture complex interactions between features, especially categorical variables, through learned embeddings. In this study, the model architecture consists of multiple transformer blocks that process embedded categorical features alongside normalized numeric features, followed by a multi-layer perceptron (MLP) classifier to generate predictions of diabetes diagnosis and disease stage. This approach allows the model to exploit non-linear feature relationships and patterns, providing a robust framework for accurate classification on large-scale clinical tabular data. The TabTransformer model architecture can be seen in Figure 2.

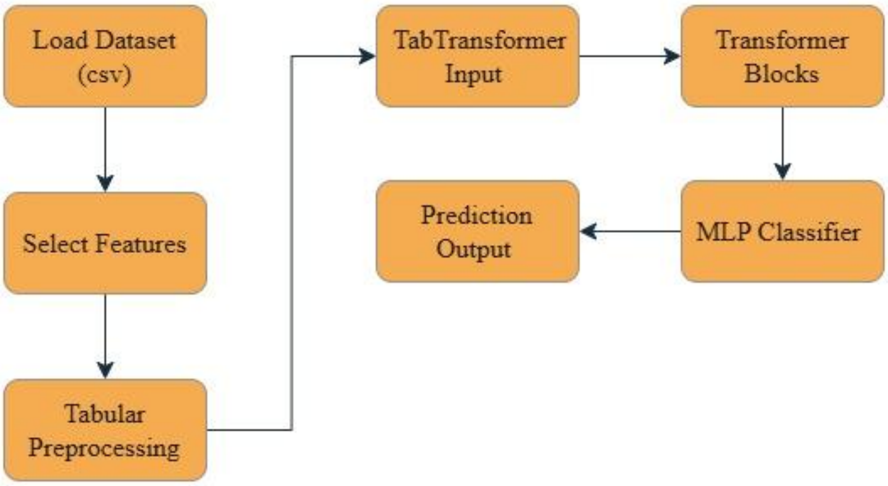


Figure 2. Architecture Of The Tab Transformer Model

Figure 2 illustrates the complete pipeline for predicting diabetes status and disease stage using the TabTransformer architecture. The data used comes from a CSV file named 'diabetes_dataset.csv' with the initial step of removing the 'patient_id' column. The selected features include five numeric features (age, BMI, fasting glucose, cholesterol, HbA1c) and four categorical features (gender, smoking status, family history of diabetes, employment status). In the tabular preprocessing stage, missing values are handled (removing or using the mean), numeric feature normalization, and categorical feature encoding into an embedding are performed. The TabTransformer input processes the numeric features directly and the categorical features with a 32-dimensional embedding. The transformer block consists of 4 layers with 8 attention heads, a feedforward dimension of 128, GELU activation, and a dropout of 0.1 to generate a contextual embedding. Classification was performed using MLP with two hidden layers [64, 32], ReLU activation, dropout 0.2, and produced binary output 'diagnosed_diabetes' and multi-class 'diabetes_stage' as the final prediction results.

4. Results and Discussion

4.1. Preprocessing Data

In this stage, data preprocessing is performed to ensure the quality and suitability of the dataset before use in the modeling process. This step includes removing identity columns and features that could potentially cause data leakage, as well as rows with no values in the target columns. After preprocessing, the number of attributes in the dataset was reduced from 31 to 27 columns without changing the number of data rows. The results of data preprocessing can be seen in Table 1.

Table 1. Data Preprocessing

Stage	Number of Rows	Number of Columns	Description
Before preprocessing	100.000	31	Raw dataset after loading process from CSV file
Setelah drop leakage & missing target	100.000	27	Several columns that could potentially cause data leaks as well as untargeted rows were deleted.

Table 1 shows the results of the data preprocessing stage performed before the modeling process. This stage removed several features that could potentially cause data leakage, as well as rows with no values in the target columns. This process reduced the number of columns from 31 to 27, while maintaining the 100,000 rows in the original dataset.

4.1 Feature Engineering

At this stage, the target and predictor variables were systematically separated. All features except the target column were placed as input (X), while the diagnosed_diabetes column was used as the output variable (y). Furthermore, the predictor features were grouped by data type to support further transformation processes. Details of the results from this stage are presented in Table 2.

Table 2.Feature Engineering

Components	Results
Target variable (y)	The diagnosed diabetes column was successfully separated as an output variable.
Predictor variable (X)	All columns other than the target are defined as inputs to the model.
Number of categorical features	There are 6 categorical type features
Number of numeric features	There are 20 features of numeric type
Readiness for the next stage	The dataset is ready to undergo further transformation processes such as encoding and normalization.

4.2 Model Results And Performance.

This section presents the results and performance evaluation of the TabTransformer model applied to the diabetes dataset. The model was trained to predict both diabetes diagnosis (binary classification) and diabetes stage (multi-class classification). Performance metrics such as accuracy, precision, recall, F1-score, and loss values are reported to assess the model's effectiveness.

Table 3. Performance Model

Model	Class	Precision	Recall	F1-score	Accuracy
TabTransformer	0	0.7716	0.8006	0.7858	0.8255
	1	0.8637	0.8420	0.8527	

Table 1 presents the performance evaluation of the TabTransformer model for predicting diabetes diagnosis. The results are reported for each class, where class 0 represents non-diabetic patients and class 1 represents diabetic patients. The model achieved an accuracy of 82.55%, demonstrating strong overall predictive capability. For class 0, the precision, recall, and F1-score were 0.7716, 0.8006, and 0.7858, respectively, indicating that the model effectively identifies non-diabetic cases. For class 1, the model achieved a higher precision of 0.8637 and recall of 0.8420, resulting in an F1-score of 0.8527, reflecting its robustness in correctly predicting diabetic cases. These results suggest that the TabTransformer model is capable of capturing complex patterns in the clinical tabular data, leading to reliable classification performance across both target classes.

http://doi.org/10.65126/jocosir.v2i2.52

4.3 ROC Curve (Receiver Operating Characteristic Curve)

The ROC (Receiver Operating Characteristic Curve) curve is used to demonstrate the model's ability to discriminate between positive and negative classes. This curve visualizes the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) at various thresholds, providing a comprehensive overview of the classification model's performance. Furthermore, the area under the curve (AUC) provides a quantitative measure of the model's discriminatory ability, with higher AUC values indicating better performance. The results can be seen in Figure 3.

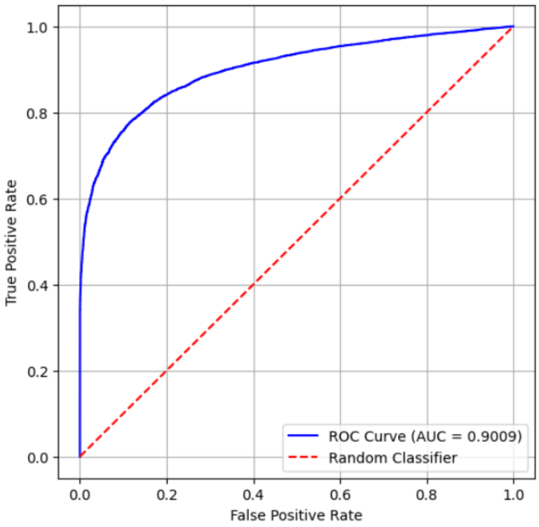
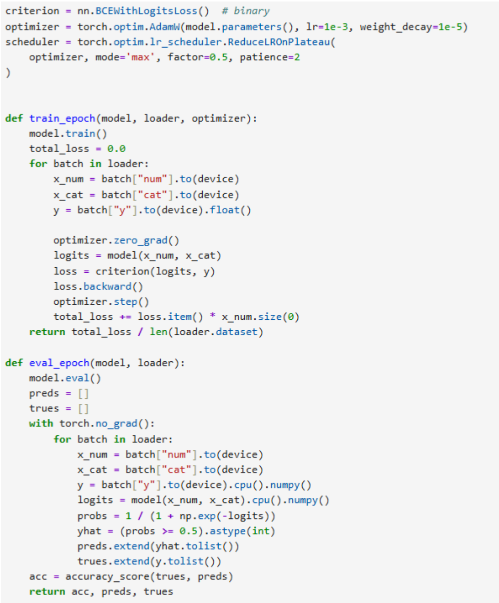


Figure 3. Receiver Operating Characteristic Curve

As shown in Figure 3, the ROC Curve illustrates the classification performance of the TabTransformer model in distinguishing patients with and without diabetes. The model achieved a high ROC AUC score of 0.9009, indicating excellent discriminative capability. This result demonstrates that the model is able to correctly identify positive cases while minimizing false positives, reflecting its robustness and reliability for predicting diabetes based on clinical tabular data.

4.4 Model Implementation

This stage implements a machine learning model using the TabTransformer architecture as the primary algorithm for classifying diabetes status. The model is initialized with predetermined parameters and then trained using data that has undergone preprocessing and feature engineering. Additionally, an optimization mechanism, loss function, and evaluation scheme are developed to ensure the training process runs systematically and produces optimal results. The implementation can be seen in Figure 4.



```
criterion = nn.BCEWithLogitsLoss()  # binary
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-3, weight_decay=1e-5)
scheduler = torch.optim.lr_scheduler.ReduceLROnPlateau(
    optimizer, mode='max', factor=0.5, patience=2
)


def train_epoch(model, loader, optimizer):
    model.train()
    total_loss = 0.0
    for batch in loader:
        x_num = batch["num"].to(device)
        x_cat = batch["cat"].to(device)
        y = batch["y"].to(device).float()

        optimizer.zero_grad()
        logits = model(x_num, x_cat)
        loss = criterion(logits, y)
        loss.backward()
        optimizer.step()
        total_loss += loss.item() * x_num.size(0)
    return total_loss / len(loader.dataset)

def eval_epoch(model, loader):
    model.eval()
    preds = []
    trues = []
    with torch.no_grad():
        for batch in loader:
            x_num = batch["num"].to(device)
            x_cat = batch["cat"].to(device)
            y = batch["y"].to(device).cpu().numpy()
            logits = model(x_num, x_cat).cpu().numpy()
            probs = 1 / (1 + np.exp(-logits))
            yhat = (probs >= 0.5).astype(int)
            preds.extend(yhat.tolist())
            trues.extend(y.tolist())
    acc = accuracy_score(trues, preds)
    return acc, preds, trues
```

Figure 4. Code Coding implementation of TabTransformer architecture

Figure 4 shows part of the TabTransformer model training process, where the BCEWithLogitsLoss loss function is defined for the binary classification case, as well as the AdamW optimizer and ReduceLROnPlateau scheduler to regulate the learning rate during training. The train_epoch() function is used to run one training cycle by performing a forward pass, loss calculation, backpropagation, and updating the model weights on each data batch. Meanwhile, the eval_epoch() function is used to evaluate the model without the weight update process, by generating probability-based predictions, performing classification using a 0.5 threshold, and calculating accuracy as a model performance metric on the validation data. Thus, these two functions act as core utilities for running the model training and evaluation process systematically.

4.5 Discussion

Model implementation results demonstrate that the TabTransformer architecture is capable of learning complex patterns in tabular clinical data containing a combination of categorical and numeric features. The self-attention mechanism in the transformer allows the model to contextually capture dependencies between categorical features before combining them with numeric features and processing them with the MLP for classification. The training phase was systematically executed using the BCEWithLogitsLoss loss function for binary classification, and the AdamW optimizer and ReduceLROnPlateau scheduler to maintain the stability of the learning process. Model evaluation was performed independently without weight updates to ensure objective performance measurements on validation data.

Overall, this implementation process resulted in a model with competitive performance for diabetes classification tasks, as reflected by high accuracy, F1-score, and ROC-AUC values. The model not only demonstrated the ability to accurately identify patients with diabetes but also maintained consistent performance in the non-diabetic class. Thus, the application of TabTransformer to large-scale medical data has proven effective in addressing tabular data-based clinical prediction problems and opens up opportunities for the development of more accurate and reliable machine learning-based medical decision support systems.

5. Conclusions

Based on the research results, the application of the TabTransformer architecture to clinical tabular data for diabetes prediction has been proven to provide robust and reliable performance. The pre-processing and feature engineering processes successfully prepare the data systematically, while the self-attention mechanism in TabTransformer is able to capture complex relationships between numerical and categorical features. The model evaluation shows an accuracy of 82.55% with a ROC AUC value of 0.9009, indicating the model's excellent discriminatory ability in distinguishing diabetic and non-diabetic patients. Thus, TabTransformer is worthy of consideration as an effective approach for the development of large-scale medical data-based prediction systems.

6. References

[1] R. Arania, T. Triwahyuni, F. Esfandiari, and F. R. Nugraha, “Hubungan antara usia, jenis kelamin, dan tingkat pendidikan dengan kejadian diabetes mellitus di Klinik Mardi Waluyo Lampung Tengah,” J. Medika Malahayati, vol. 5, no. 3, pp. 146–153, 2021.

[2] U. P. Sukmara, “Meningkatkan kesadaran pencegahan penyakit tidak menular pada hipertensi dan diabetes melitus melalui edukasi di masyarakat,” J. Medika: Medika, vol. 4, no. 3, pp. 436–441, 2025.

[3] M. R. Febriansyah, “Diplomasi Kesehatan International Diabetes Federation (IDF) untuk meningkatkan kesadaran diabetes di kawasan Pasifik Barat pada tahun 2017–2022,” Doctoral Dissertation, Universitas Islam Indonesia, 2024.

[4] H. Sun et al., “IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045,” Diabetes Res. Clin. Pract., vol. 183, p. 109119, 2022.

[5] P. Saeedi et al., “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition,” Diabetes Res. Clin. Pract., vol. 157, p. 107843, 2019.

[6] I. Irwansyah and I. S. Kasim, “Deteksi Dini Risiko Diabetes Melitus pada Staff Pengajar Stikes Megarezky Makassar,” J. Ilm. Kesehatan Sandi Husada, vol. 9, no. 1, pp. 540–547, 2020.

[7] F. K. Adli, “Diabetes Melitus Gestasional: Diagnosis dan Faktor Risiko,” J. Medika Hutama, vol. 3, no. 01 Oktober, pp. 1545–1551, 2021.

[8] H. Siswanti, F. A. Rohmaniah, S. Sukarmin, and M. Ridwanto, “Deteksi dini faktor risiko sebagai upaya pencegahan penyakit Diabetes Mellitus,” J. Inov. Penelitian dan Pengabdian Masyarakat, vol. 5, no. 1, pp. 118–127, 2025.

[9] V. Agustina et al., “Deteksi dini penyakit diabetes melitus,” Magistrorum et Scholarium: J. Pengabdian Masyarakat, vol. 2, no. 2, pp. 300–309, 2021.

[10] H. Purnama, H. Z. N. Adzidzah, M. Solihat, and M. Septriani, “Determinan risiko dan pencegahan terhadap kejadian penyakit diabetes melitus tipe 2 pada usia produktif di Wilayah DKI Jakarta,” J. Public Health Educ., vol. 2, no. 4, pp. 158–166, 2023.

[11] A. Khaidar, M. Arhami, and M. Abdi, “Application of the Random Forest Method for UKT Classification at Politeknik Negeri Lhokseumawe,” J. Artif. Intell. Softw. Eng., vol. 4, no. 2, pp. 94–103, 2024.

[12] A. Al Khaidar, “Analisis sentimen di Instagram terhadap Menteri Keuangan Purbaya Yudhi Sadewa menggunakan metode Logistic Regression,” J. Inform. dan Teknik Elektro Terapan, vol. 13, no. 3S1, 2025.

[13] I. N. Migdalis, “Chronic Complications of Diabetes: Prevalence, Prevention, and Management,” Journal of Clinical Medicine, vol. 13, no. 23, p. 7001, 2024, doi: 10.3390/jcm13237001.

[14] D. Tomic et al., “The burden and risks of emerging complications of diabetes mellitus,” 2022, doi: 10.1177/.... (detail jurnal tidak dicantumkan pada sumber).

[15] S. A. Antar, “Diabetes mellitus: Classification, mediators, and emerging therapies,” 2023, doi: 10.1016/S0753-3322(23)01532-9.

[16] X. Lu et al., “Type 2 Diabetes Mellitus in adults: Pathogenesis, complications and management,” Signal Transduction and Targeted Therapy, 2024, doi: 10.1038/s41392-024-01951-9.

[17] “Global, regional, and national burden of type 2 diabetes mellitus from 1990 to 2021, with projections to 2050,” Frontiers in Endocrinology, 2024, doi: 10.3389/fendo.2024.1501690.

[18] M. Yoon, J. Park, and S. Lee, “Transformer-based deep learning model for predicting chronic disease progression using longitudinal health records,” *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 9, pp. 4321–4332, 2023, doi: 10.1109/JBHI.2023.3265512.

[19] A. Gupta, R. Das, and S. K. Singh, “Attention-based deep neural networks for diabetes diagnosis using clinical tabular datasets,” *Computers in Biology and Medicine*, vol. 158, p. 106993, 2023, doi: 10.1016/j.compbiomed.2023.106993.

[20] M. Rahman et al., “Deep learning for diabetes prediction using structured electronic health records,” BMC Medical Informatics and Decision Making, vol. 23, no. 1, 2023, doi: 10.1186/s12911-023-02291-3.

[21] H. Wang, F. Wu, and Y. Zhou, “Transformer-based model for tabular clinical data classification,” IEEE J. Biomed. Health Inform., vol. 28, no. 1, pp. 112–122, 2024.

[22] D. Wang, H. Lin, and F. Zhou, “A transformer-driven framework for risk prediction using structured electronic health records,” *Applied Intelligence*, vol. 53, no. 18, pp. 21547–21562, 2023, doi: 10.1007/s10489-023-04655-1.

[23] Y. Li and M. Zhang, “Prediction of diabetes using deep neural networks on real-world clinical datasets,” Computers in Biology and Medicine, vol. 160, p. 107216, 2023.

[24] M. Sun et al., “Medical tabular data classification using hybrid attention network,” Applied Intelligence, vol. 53, pp. 15230–15244, 2023.

[25] R. Kaur and P. Singh, “An improved transformer-based model for diabetes prediction,” IEEE Access, vol. 11, pp. 44512–44522, 2023.

[26] J. Chen, X. Liu, and Y. Gao, “Attention-guided embedding learning for clinical tabular prediction,” Expert Systems with Applications, vol. 238, p. 121801, 2024.

[27] L. Patel et al., “Artificial intelligence in diabetes diagnosis and risk stratification,” Frontiers in Endocrinology, vol. 14, pp. 1–15, 2023, doi: 10.3389/fendo.2023.1180922.

[28] S. Ö. Arık and T. Pfister, “TabTransformer: Tabular data modeling using contextual embeddings,” in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 8, pp. 6679–6687, 2021.

[29] S. Huang, Z. Chen, J. Li, and K. Ma, “TabTransformer for electronic health record classification,” IEEE Access, vol. 10, pp. 125834–125846, 2022, doi: 10.1109/ACCESS.2022.3220114.

[30] S. Xu, Q. Zhang, and H. Liu, “Self-attention based deep models for chronic disease risk prediction using structured EHR,” Int. J. Med. Inform., vol. 177, p. 105237, 2023, doi: 10.1016/j.ijmedinf.2023.105237.