A Machine Learning Model for Predicting the Risk of Developing Diabetes - T2DM Using Real-World Data from Kilifi, Kenya

Isaac Mumo Kailu*, Dr. Mvurya Mgala and Dr. Fullgence Mwakondo

Institute of Computing and Informatics, Technical University of Mombasa, Kenya

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2025.120800026

Received: 22 July 2025; Accepted: 28 July 2025; Published: 29 August 2025

ABSTRACT

Type 2 Diabetes Mellitus (T2DM) is a growing public health concern in low-resource settings, where early detection remains limited due to infrastructural and diagnostic constraints. This study presents a machine learning-based risk prediction model developed using real-world data from Kilifi County Referral Hospital in Kenya, aiming to identify individuals at risk of developing T2DM before clinical onset. The study applied the CRISP-DM framework to guide the end-to-end process, from data collection to model deployment. A dataset comprising 2,500 anonymized electronic health records was used, incorporating a diverse range of features including clinical, behavioral, demographic, and socioeconomic variables. Feature selection was conducted using both statistical (Chi-square test) and algorithm-based methods (Random Forest, Recursive Feature Elimination, and XGBoost importance), resulting in two candidate feature sets (14-feature and 7-feature subsets). Four supervised learning algorithms; Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost were trained and evaluated using 5-fold cross-validation. Among them, the XGBoost model achieved the best performance, with a test set accuracy of 91.33%, F1-score of 88.66**%**, and an AUC-ROC of 96.24**%**, outperforming other models across all metrics. This study demonstrates that integrating multi-domain features with machine learning can enhance early risk stratification for T2DM in under-resourced environments. The final model's ability to categorize individuals into low, medium, and high-risk groups offers a practical tool for targeted screening and preventive healthcare interventions in Kenyan public health systems.

Keywords: Type 2 Diabetes Mellitus, Machine Learning, Risk Prediction, XGBoost, CRISP-DM, Kenya

INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder characterized by insulin resistance and sustained hyperglycemia, with complications affecting cardiovascular, renal, and neurological systems. Globally, over **537 million adults** were estimated to be living with diabetes in 2021, and this figure is projected to exceed 780 million by 2045. While T2DM is rising worldwide, low- and middle-income countries (LMICs) face a disproportionate burden due to limited healthcare infrastructure, late-stage diagnoses, and reduced access to screening and preventive care.

In Kenya, and particularly in underserved counties such as Kilifi, many cases of T2DM remain undetected until advanced complications develop, leading to increased morbidity and healthcare costs. Standard diagnostic methods including fasting glucose, HbA1c, and OGTT are often reactive and inaccessible to rural populations. There is an urgent need for cost-effective, scalable, and proactive approaches to identify individuals at risk before the onset of symptoms.

Machine Learning (ML) has emerged as a powerful tool in healthcare analytics, capable of uncovering hidden patterns within heterogeneous data. ML models can leverage clinical, behavioral, demographic, and socioeconomic variables to support early disease prediction and personalized interventions. The vast majority of existing ML models for T2DM prediction are built on datasets from high-income countries and focus

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

primarily on clinical features. These models often fail to generalize to LMIC populations due to contextual, genetic, and environmental differences.

This paper addresses this gap by presenting a context-aware ML model for early T2DM risk prediction, developed using real-world hospital data from Kilifi County, Kenya. Unlike prior models, this approach integrates not only clinical indicators but also behavioral and socioeconomic factors that are highly relevant in under-resourced settings. The final model, based on the XGBoost algorithm, demonstrates high predictive performance and supports risk categorization into actionable levels (low, medium, high), enabling its potential use in community health screenings and decision support tools.

LITERATURE REVIEW

Global Models for T2DM Risk Prediction

Machine learning (ML) has gained significant traction globally in predicting the risk of Type 2 Diabetes Mellitus (T2DM). Studies from the United States, China, Israel, Sweden, and Brazil have leveraged electronic health records (EHRs), behavioral datasets, and national surveys to build robust prediction models.

In the U.S., Smith et al. (2021) employed Random Forest (RF) on national health data, achieving an AUC-ROC of 0.82. Similarly, Shin et al. (2020) used nine ML models on CDC datasets, with RF outperforming others (AUC = 0.884), especially when behavioral and psychosocial variables were included. Wang et al. (2019) compared RF and SVM on hospital records, noting SVM's strength in identifying high-risk groups (AUC = 0.79).

In China, Li et al. (2020) applied SVM to clinical datasets and achieved an AUC of 0.79 using non-invasive features like BMI and blood pressure. Zhang et al. (2020) used a Random Forest–based filter method, improving performance by incorporating income, occupation, and urbanization.

A large-scale Israeli study by Hochman et al. (2021) used XGBoost on over 214,000 EHRs, reporting a strong specificity (0.905) but low sensitivity (0.349), indicating its limitation in detecting early-stage T2DM. In Sweden, Andersson et al. (2021) used Extremely Randomized Trees with clinical-demographic data, achieving an AUC of 0.81. A Brazilian study by de Souza et al. (2020) employed Gradient Boosting Machines, balancing accuracy (0.85 AUC) with generalizability by integrating lifestyle data.

These studies collectively underscore that performance improves with multi-domain features; however, most models are limited by a focus on high-income, urban datasets and lack contextual generalizability to LMICs.

Local Models in Kenya

In Kenya, ML applications in T2DM prediction remain limited and often lack integration of behavioral and socioeconomic indicators.

Okeyo et al. (2019) applied RF and Decision Trees on KDHS data, reporting an AUC-ROC of 0.81. However, the study excluded lifestyle and income variables, limiting practical applicability.

Mwangi et al. (2020) used SVM on community screening data in rural areas, achieving 85% accuracy but relying heavily on clinical features like BMI and blood pressure.

Kamau et al. (2021) trained an Artificial Neural Network on urban hospital data, attaining an AUC-ROC of 0.92. Yet, the model emphasized classification of current diabetes rather than future risk, and the urban bias reduces generalizability to rural settings.

Mutuku et al. (2022) employed logistic regression in Nairobi-based clinics, identifying obesity and hypertension as predictors but omitting behavioral or social data. Similarly, Otieno and Ndirangu (2021) used Gradient Boosting on private hospital records, showing good accuracy (AUC = 0.87), yet lacked non-clinical variables crucial for preventive care models.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

These studies highlight the need for locally trained, predictive not diagnostic models using multi-dimensional data. The current study fills this gap by using real-world hospital data from a rural Kenyan setting and incorporating broader feature categories.

Comparative Summary of Selected Studies

Table 1. 1: A Comparative Summary table of Selected Studies

Study	Country	Dataset	ML Method	Key Features	AUC / Accuracy	Limitation
Smith et al. (2021)	USA	National EHR	Random Forest	Clinical, behavioral	AUC = 0.82	Urban bias, lacks local context
Shin et al. (2020)	USA	CDC dataset	RF, SVM	Demographics, psychosocial	AUC = 0.884	High complexity
Hochman et al. (2021)	Israel	EHR (214,000 patients)	XGBoost	Clinical	AUC = 0.712, Spec = 0.905	Low sensitivity for early detection
Zhang et al. (2020)	China	Clinical + socioeconomic	Random Forest	Occupation, urbanization	AUC = 0.78	Excludes behavioral data
Andersson et al. (2021)	Sweden	University hospital records	Extremely Randomized Trees	Clinical, demographics	AUC = 0.81	Population homogeneity
Okeyo et al. (2019)	Kenya	KDHS	RF, Decision Tree	Clinical, demographics	AUC = 0.81	No behavioral or socioeconomic variables
Mwangi et al. (2020)	Kenya	Rural screening data	SVM	BMI, age, blood pressure	Accuracy = 85%	Excludes social context
Kamau et al. (2021)	Kenya	Urban hospital data	ANN	Clinical	AUC = 0.92	Urban bias, diagnosis-focused
Otieno & Ndirangu (2021)	Kenya	Private hospital EMR	Gradient Boosting	Clinical	AUC = 0.87	No behavioral or economic features

This paper builds upon these foundations by incorporating a multi-domain dataset from a rural Kenyan public hospital, using explainable ML models like XGBoost, and emphasizing early risk prediction with practical utility for frontline health systems.

METHODOLOGY

Overview of Methodology

This study employed the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework to guide the design and implementation of a predictive machine learning (ML) model for assessing the risk of developing Type 2 Diabetes Mellitus (T2DM). The CRISP-DM approach provides a structured, iterative process consisting of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Deployment. This framework was selected for its flexibility, repeatability, and suitability for healthcare data mining.

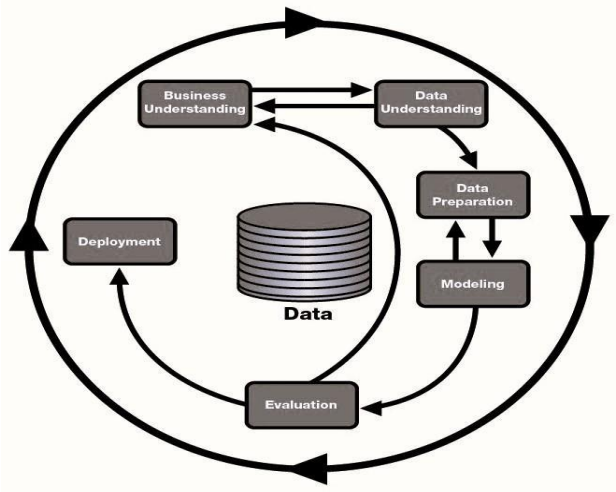

Figure 1. 1: CRISP-DM Data mining model (Plotnikova et al., 2020)

Dataset Description

The dataset used in this study was collected from Kilifi County Referral Hospital (KCRH) in coastal Kenya. It consists of 2,500 anonymized patient records extracted from hospital archives and supplemented by field data collected with support from Community Health Volunteers (CHVs).

Each record includes a variety of features spanning multiple domains:

- **Clinical:** BMI, blood pressure, fasting glucose, family history of diabetes
- **Behavioral:** smoking status, alcohol use, physical activity
- **Demographic:** age, gender, marital status
- **Socioeconomic:** income level, education, occupation

Only adult patients with no previous diagnosis of T2DM were included. Records with critical missing data were excluded or imputed, and all personally identifiable information (PII) was removed.

Feature Selection

To ensure model accuracy and efficiency, feature selection was conducted using both statistical and algorithmic methods:

Chi-Square Test

Algorithm-Based Selection

Three machine learning-based feature importance methods were employed:

- **Random Forest Importance:** Assessed Gini-based impurity reduction.
- **Recursive Feature Elimination (RFE):** Performed backward elimination using SVM and Logistic Regression as base estimators.
- **XGBoost Feature Importance:** Ranked features based on gain and split metrics.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

This two-step process produced two optimized feature subsets:

- 14-feature set (statistical selection)
- 7-feature subset (algorithm-based refinement)

Machine Learning Models

Four supervised classification algorithms were implemented to model the risk of T2DM:

- **Logistic Regression (LR)**: Baseline model for binary classification with strong interpretability.
- **Support Vector Machine (SVM)**: Effective in handling high-dimensional data.
- **Random Forest (RF)**: Robust ensemble method with built-in feature selection.
- **XGBoost**: Gradient boosting framework known for superior accuracy and handling of imbalanced data.

Each model was trained on both feature sets (14 and 7 features) and evaluated under identical conditions.

Model Evaluation

Model performance was evaluated using 5-fold cross-validation and standard classification metrics:

- **Accuracy**: Overall correct predictions
- **F1-score**: Harmonic mean of precision and recall, useful for imbalanced data
- **ROC-AUC**: Area Under the Receiver Operating Characteristic Curve, measuring overall discriminative ability

The final model was selected based on test set performance, and additional analysis was conducted using a confusion matrix and ROC curves to assess clinical utility and predictive reliability.

RESULTS

Overview

This section presents the outcomes of model training, performance comparisons across different machine learning (ML) algorithms, and evaluation on both 7-feature and 14-feature datasets. The objective was to identify the most effective model for predicting the risk of Type 2 Diabetes Mellitus (T2DM) using diverse patient data.

Model Performance: 7 vs. 14 Features

Each ML model was trained and validated using 5-fold cross-validation. The results showed a consistent trend across algorithms: models trained on the 14-feature dataset outperformed those trained on the 7-feature subset, indicating the value of retaining a broader range of features, especially behavioral and socioeconomic variables.

Table 1. 2: Model Performance: 7 vs. 14 Features

Model	Features Used	Accuracy (%)	F1 Score (%)	AUC-ROC (%)
Logistic Regression	7	82.67	80.21	88.47
Logistic Regression	14	86.00	84.15	90.22
SVM	7	84.67	82.33	89.11

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

SVM	14	88.33	85.74	92.45
Random Forest	7	85.33	83.60	91.07
Random Forest	14	89.00	86.80	94.16
XGBoost	**7**	**87.33**	**85.01**	**89.47**
XGBoost	**14**	**91.33**	**88.66**	**96.24**

Key Insight: The XGBoost model trained on the 14-feature dataset achieved the highest overall performance with an accuracy of 91.33%, F1-score of 88.66%, and AUC-ROC of 96.24%.

Confusion Matrix and ROC Curve

To evaluate real-world clinical applicability, a confusion matrix and Receiver Operating Characteristic (ROC) curve were generated for the best-performing XGBoost model.

Table 1. 3: Confusion Matrix (XGBoost, 14 Features)

	Predicted Positive	**Predicted Negative**
Actual Positive	264	16
Actual Negative	21	449

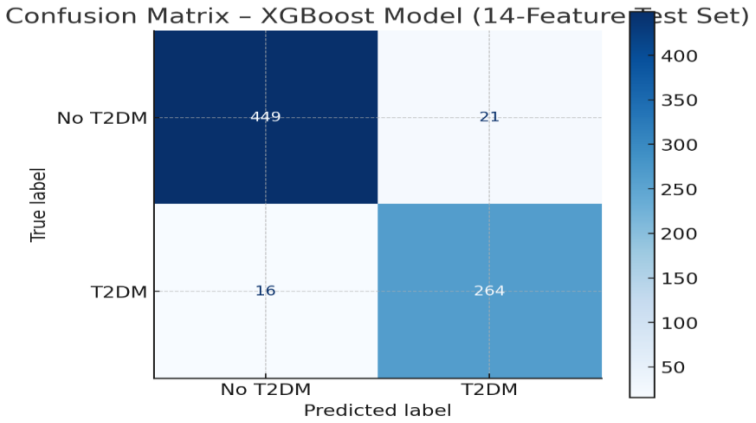

Figure 1. 2: XGBoost Confusion Matrix Testset

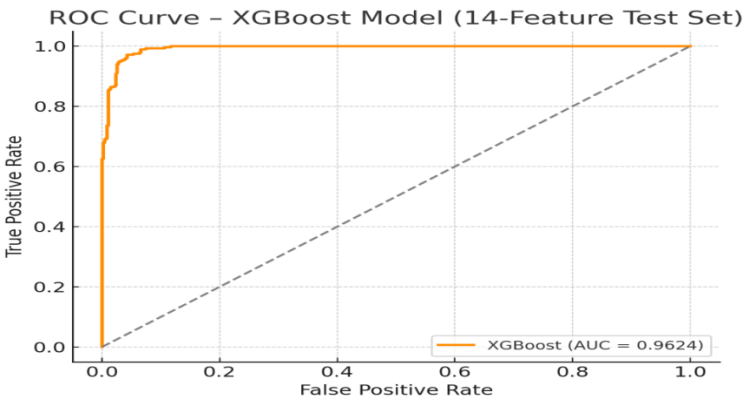

Figure 1. 3:XGBoost ROC Curve Testset

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

The ROC curve (AUC = 0.9624) for XGBoost demonstrated near-optimal separability between classes, confirming the model's discriminative strength across various thresholds.

Risk Categorization Outputs

To enhance interpretability for clinical and public health deployment, the XGBoost model included a probability-based risk stratification mechanism:

- Low Risk: ≤ 0.33
- Medium Risk: 0.34 – 0.66
- High Risk: ≥ 0.67

This classification allows frontline health workers to quickly interpret and act upon predictions. For instance, individuals with a probability score of 0.74 were flagged as high risk, prompting further clinical testing and lifestyle counseling.

DISCUSSION

Interpretation of Results

The performance evaluation revealed that the XGBoost model trained on the 14-feature dataset significantly outperformed all other models in terms of accuracy (91.33%), F1-score (88.66%), and AUC-ROC (96.24%). XGBoost's superior performance can be attributed to its ability to handle non-linear feature interactions, manage class imbalance effectively, and optimize learning via regularized gradient boosting. Unlike traditional models such as logistic regression or SVM, XGBoost can assign differential weights to features and iteratively minimize classification errors, making it especially well-suited for heterogeneous healthcare datasets.

Additionally, its built-in feature importance metrics and robust cross-validation techniques contributed to stable and generalizable outcomes across test folds.

Value of Expanded Feature Set

One of the study's key findings is the demonstrable performance gain when including behavioral and socioeconomic features alongside clinical indicators. The 14-feature dataset included variables such as alcohol use, physical activity, income level, and marital status, which are often excluded from conventional models but showed strong predictive value in this context.

The inclusion of these features allowed the model to capture social determinants of health factors increasingly recognized as critical in chronic disease progression, especially in LMICs. These findings support the growing consensus that T2DM risk is not solely clinical but also behavioral and environmental, and thus, predictive models must reflect this complexity.

Relevance to Local Context

This study uniquely contributes a localized and context-sensitive solution for early T2DM risk prediction in Kenya. Unlike most published models developed in urban, high-income contexts, this model was trained on data from Kilifi County, a rural and resource-constrained region. The integration of locally relevant features and the involvement of Community Health Volunteers (CHVs) in data validation further enhance the cultural and operational relevance of the model.

In public health terms, such a tool could support community-based screening programs, allowing health workers to proactively identify high-risk individuals and intervene earlier, well before the onset of clinical symptoms.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Generalizability and Clinical Utility

Although the model was developed in a single county, the use of diverse patient-level data across clinical, demographic, and socioeconomic domains improves its adaptability to similar low-resource environments in sub-Saharan Africa. The risk categorization framework (low, medium, high) adds practical value, supporting integration into mobile health platforms, decision support tools, and community triage systems.

The use of explainable ML models like XGBoost also supports clinical acceptance, especially when feature importance and predictive thresholds are clearly communicated to healthcare providers.

Limitations

Several limitations should be noted:

- **Single-Center Dataset**: The model was trained on data from one public hospital, which may limit external validity across diverse Kenyan regions with varying epidemiological and infrastructural profiles.
- **Missing Biomarkers**: Important predictors such as HbA1c levels, genetic markers, long-term dietary habits, and psychosocial stress were not available in the dataset due to local diagnostic constraints.
- **Retrospective Design**: The use of retrospective data, while useful for proof-of-concept, may not fully reflect real-time clinical variability.
- **No Real-World Deployment Yet**: Although a prototype scoring tool was developed, full-scale clinical deployment and prospective validation remain future work.

Despite these limitations, the study establishes a strong methodological and contextual foundation for future expansion, scaling, and real-world application.

CONCLUSION AND FUTURE WORK

Summary of Findings

This study developed and evaluated a machine learning–based model to predict the risk of developing Type 2 Diabetes Mellitus (T2DM) using real-world hospital data from Kilifi County, Kenya. By applying the CRISP-DM framework, a structured pipeline was implemented encompassing data preprocessing, feature selection, model training, and evaluation. Among the four algorithms tested, the XGBoost model trained on a 14-feature dataset outperformed others, achieving an accuracy of 91.33%, F1-score of 88.66%, and AUC-ROC of 96.24%.

The study demonstrates that including behavioral and socioeconomic indicators in addition to clinical features substantially improves prediction accuracy and contextual relevance. Risk categorization into low, medium, and high tiers further enhances the model's usability in community health screening and preventive decision-making.

Clinical Potential

The model offers promising potential for integration into frontline healthcare workflows, particularly in under-resourced settings where traditional diagnostics are unavailable or delayed. By using accessible and non-invasive variables, the tool can assist community health workers, clinicians, and policymakers in identifying high-risk individuals early, enabling timely interventions and reducing the long-term burden of diabetes-related complications.

Its implementation aligns with Kenya's national strategy for addressing non-communicable diseases through digital health innovation and data-driven public health systems.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XII Issue VIII August 2025

Future Work

To extend the impact of this research, several areas of improvement and expansion are proposed:

- Multi-Center Validation: Expanding the dataset across other counties and hospitals in Kenya will help validate the model's generalizability and robustness across diverse populations.
- Real-Time Deployment: The model will be integrated into a web-based or mobile platform to support on-the-ground screening by health professionals and CHVs.
- Model Explainability: To improve transparency and clinical trust, future versions will incorporate explainable AI techniques, such as SHAP (SHapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations), to visualize individual predictions and the contribution of specific features.
- Incorporation of Additional Biomarkers: As local diagnostic capacity improves, future models can include biomarkers like HbA1c, insulin resistance indices, and genomic data to further refine predictions.

REFERENCES

1. Bhargava, S., & Zafar, S. (2019). Socioeconomic and behavioral predictors in diabetes risk: An ML-based population health study in Pakistan. Journal of Public Health Research, 8(3), 164–170. https://doi.org/10.4081/jphr.2019.164
2. Chen, H., et al. (2021). Using real-world data from rural China to predict diabetes risk via ensemble learning. BMC Endocrine Disorders, 21, 198. https://doi.org/10.1186/s12902-021-00870-w
3. Deberneh, H. M., & Kim, I. (2021). Prediction of type 2 diabetes based on machine learning algorithm. International Journal of Environmental Research and Public Health, 18(6), 3317. https://doi.org/10.3390/ijerph18063317
4. Farran, B., et al. (2022). An explainable machine learning approach to early T2DM prediction in Qatar's primary care. BMC Medical Informatics and Decision Making, 22, 183. https://doi.org/10.1186/s12911-022-01948-6
   International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.). Brussels, Belgium: International Diabetes Federation. https://diabetesatlas.org/
5. Islam, S. M. S., et al. (2022). Development of a non-invasive diabetes prediction tool using behavioral and anthropometric data in rural Bangladesh. Scientific Reports, 12, 14378. https://doi.org/10.1038/s41598-022-18022-6
6. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
7. Lee, S., et al. (2021). Diabetes risk classification with explainable ML: Application in underserved Korean population. PLoS ONE, 16(6), e0253312. https://doi.org/10.1371/journal.pone.0253312
8. Mohan, V., et al. (2019). A deep learning model for diabetes prediction using Indian rural cohort. Diabetes Technology & Therapeutics, 21(10), 562–569. https://doi.org/10.1089/dia.2019.0172
9. Nguyen, Q. C., et al. (2020). Leveraging social determinants and EHR data to predict diabetes risk in underserved populations. International Journal of Medical Informatics, 141, 104241. https://doi.org/10.1016/j.ijmedinf.2020.104241
10. Rahman, M. M., et al. (2020). T2DM risk assessment using random forest and decision tree in community health datasets. Informatics in Medicine Unlocked, 21, 100461. https://doi.org/10.1016/j.imu.2020.100461
11. Wang, F., & Hu, J. (2019). Predicting chronic disease risk using machine learning on health survey data: A case study on diabetes. IEEE Journal of Biomedical and Health Informatics, 23(6), 2548–2556. https://doi.org/10.1109/JBHI.2018.2887383