Analysis of Datasets Created to Assess the Risk of Developing Gestational Diabetes Mellitus*

Mukhriddin Arabboev[1,*,†], Shohruh Begmatov[1,†], Mokhirjon Rikhsivoev[1,†], Saidakmal Saydiakbarov[1,†], Zukhriddin Khamidjonov[1,†], Sardor Vakhkhobov[1,†], Khurshid Aliyarov[1,†] and Khabibullo Nosirov[1,†]

[1] *Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, 108 Amir Temur St., Tashkent, 100084, Uzbekistan*

Abstract

In recent years, the healthcare field has seen a rise in the use of artificial intelligence. There is growing interest in applying artificial intelligence technology to the field of healthcare. To effectively predict disease and deploy proper artificial intelligence and machine learning algorithms, there is a need for suitable datasets. Datasets are widely used to assess the risk of developing diabetes, one of the most common diseases. Given the preceding, this paper reviews datasets created to assess the risk of developing gestational diabetes mellitus (GDM) used worldwide.

Keywords

Dataset, gestational diabetes mellitus, machine learning

1. Introduction

In recent years, the number of people with diabetes has been increasing worldwide. Diabetes is one of the most common diseases among the population [1]. There are common types of diabetes such as type 1 [2], type 2 [3] and gestational diabetes [4]. Gestational Diabetes Mellitus (GDM) poses significant health risks to both mothers and infants, making its early detection and effective management crucial for maternal and fetal well-being. As the prevalence of GDM continues to rise globally, there is a growing need for robust predictive models to identify women at risk. Key to developing such models is the availability and analysis of high-quality datasets specifically tailored for assessing GDM risk factors.

Artificial intelligence (AI) is a rapidly developing field, and its application in treating diabetes may revolutionize the approach to diagnosing and managing this chronic condition [5].

Machine learning algorithms are used to support predictive models for the risk of developing diabetes or its complications [6]. Digital therapy has proven to be an established intervention for lifestyle therapy in the treatment of diabetes. Patients are increasingly empowered to self-manage their diabetes, and both patients and healthcare professionals benefit from clinical decision support. AI enables continuous and remote monitoring of patient symptoms and biomarkers. Technological advances have helped optimize resource utilization in diabetes. Artificial intelligence is changing massive processes in diabetes care, from traditional treatment strategies to creating targeted, data-driven precision care.

Our review provides a comprehensive resource for researchers, IT companies involved in developing medical data, and technology companies specializing in the healthcare sector.

* *IVUS2024: Information Society and University Studies 2024, May 17, Kaunas, Lithuania*

[1,*] Corresponding author

[†] These author contributed equally.

mukhriddin.9207@gmail.com (M. Arabboev); bek.shohruh@gmail.com (Sh.Begmatov); mrikhsivoev@gmail.com (M.Rikhsivoev); saidakmalflash@gmail.com (S. Saydiakbarov); hamidjanovzuhriddin22@gmail.com (Z. Khamidjonov); sardorakbarivich@gmail.com; (S. Vakhkhobov); uzregxurshid@gmail.com (K. Aliyarov); n.khabibullo1990@gmail.com (K. Nosirov).

0000-0001-5733-5889 (M. Arabboev); 0000-0002-2441-916X (Sh.Begmatov); 0000-0002-4691-1470 (M.Rikhsivoev); 0009-0005-6654-2851 (S. Saydiakbarov).

The contributions of this survey are summarised as follows:

1. Research conducted on creating a dataset for GDM across geographic regions is presented.

2. We present a comprehensive review of research on creating a dataset for assessing the risk of gestational diabetes mellitus (GDM). This review covers the algorithms or models used, datasets used or created, and the results achieved.

3. The dependence of the number of data in the dataset on the accuracy of the model is critically analyzed.

The rest of the paper follows this structure: The section "Analysis of studies on creating datasets for gestational diabetes" reviewed up-to-date research done in the related field. The section "Results obtained in the analyzed studies" compared the results gained in the analyzed studies. The section "Conclusion" concludes this paper.

2. Analysis of studies on creating datasets for gestational diabetes

This section reviews studies on creating datasets for gestational diabetes. In preparation for this review paper, between January and February 2024, Google Scholar, PubMed, Science Direct, and IEEE Xplore databases were searched for articles with the keywords "gestational, diabetes mellitus, dataset, Artificial Intelligence". 2557 articles were found as a result of the search, of which 15 papers met specific inclusion.

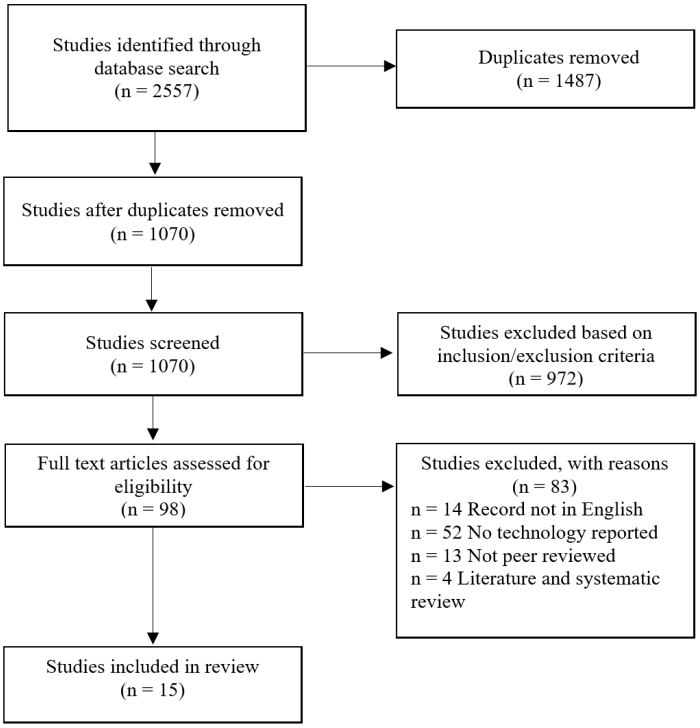

Figure 1: PRISMA Diagram of article review and selection.

In [7], it is proposed an ensemble prediction model for the diagnosis of gestational diabetes. The data collection was obtained from the laboratories of the Kurdistan region, which collected data from pregnant women with and without diabetes. The proposed model uses the KMeans clustering method for data reduction, and the elbow method to find the optimal k value and the Mahalanobis distance method to find the cluster most related to new samples. In the study, it is used classification methods such as decision tree (DT), random forest (RF), SVM, KNN, logistic regression (LR), and Naïve Bayes (NB).

In [8], a system is proposed to solve the problem of dividing diabetic patients into two categories: diabetic patients with acute illnesses and diabetic patients without acute illnesses. This study is based on the Electronic Health Records (EHR) of Osakidetza (Basque Health Service). Analytical and clinical parameter data for this study were obtained from the PREST database.

In [9], it is created a dataset using CERNER records of pregnancies observed at St Mary's Hospital, London between April 2016 and November 2019. In this study, the researchers conducted a

retrospective observational study. The initial search identified 26,063 patients with the following factors: postcode, height, weight, BMI at booking, ethnicity (self-reported), parity, glucose tolerance test offer, test results (0 min and up to 120 min after 75 g glucose load), mode of delivery, estimated total blood loss, gestational age, newborn weight, SCBU admission, length of postpartum stay, fetal sex, and stillbirth are some other factors to consider.

In [10], it is aimed at improving the diagnosis of gestational diabetes by using data collection methods. Also, this study analyzed the performance of supervised learning algorithms such as ID3, Naïve Bayes, C4.5, and Random Tree. The results of the experiment showed that the Random Tree algorithm gave the best result with the highest accuracy and the lowest error rate. The dataset used in this study is a clinical dataset collected from St. Isabella's Hospital, Mylapore, Chennai, and the National Institute of Diabetes, Digestive, and Kidney Diseases, which includes records of about 600 patients. In particular, all patients listed in the dataset were pregnant and over 21 years old.

In [11], it is developed a machine learning-based prediction model for gestational diabetes (GDM) in early pregnancy in Chinese women. This study used population-based data from 19,331 pregnant women registered as pregnant up to 15 weeks of gestation between October 2010 and August 2012 in Tianjin, China. The dataset is randomly divided into a training set (70%) and a testing set (30%). Risk factors collected during enrollment were reviewed and used to build a predictive model on the training data set. Machine learning, such as the extreme gradient boosting (XGBoost) method, was used to develop the model.

In [12], it is created a dataset based on data from 6822 pregnant women living in a geographic area defined by three regional health boards in New Zealand. The prevalence of GDM was estimated using four commonly used data sources. Coded clinical data on diabetes status were collected from regional health boards and the Ministry of Health's National Minimum Data Set, and plasma glucose results were collected from laboratories serving the recruitment area and were coded according to the New Zealand Society for the Study of Diabetes diagnostic criteria and collected via self-administered diabetes status questionnaires.

In another interesting study [13], the authors created a dataset based on data collected at the West China Second Hospital in Chengdu, Sichuan. A total of 33,935 pregnant women were enrolled in the EHR from 2013 to 2016 for experimental data. The GDM-related data of these samples contained 106 features of archival data, 23 features of audit data, 157 features of laboratory information system test data, and 268 features of EHR first pages. After data cleaning, the authors used a filtering strategy to preselect patients whose EHR data were associated with GDM as candidate samples, excluding pre-gestational diabetes. Through this process, the authors obtained an accurate data set of 10,105 samples with common clinical characteristics. This dataset contains 1649 GDM (positive) cases and 8456 non-GDM (negative) cases.

A new dataset for gestational diabetes is created in [14]. Data used in this study were collected from local hospitals in Mysuru, Karnataka, India. Medical records were obtained after anonymizing patients to ensure confidentiality. The dataset was developed by keeping obstetrics and gynecology consultants in feedback. The dataset contains information on 1352 pregnant women. The GDM dataset was developed with the help of physicians by removing less significant and irrelevant features, followed by data cleaning and transformation.

In [15], it is created a dataset based on data from the general ward of Kurmitola General Hospital in Bangladesh to test an ML model and predict diabetes for Bangladeshi patients. The authors addressed the group of trainee doctors who participated in the data collection process. They conducted a brief interview with the patients, and after their appropriate consent, the doctors agreed to provide relevant information to the authors. In total, it took about three weeks in November 2019 to collect all relevant information. This split dataset contains data from 181 patients and consists of 4 characteristics: patient age, body mass index, number of pregnancies, and glucose concentration.

In [16], the authors created a dataset based on real pregnancy test data from a hospital in Beijing from 2008 to 2018. The dataset contains examination records of 120,396 pregnant women. In the entire sample set, 18,400 pregnant women had gestational diabetes, accounting for approximately 15.28%, and 7,518 pregnant women had gestational hypertension, accounting for approximately 6.24%.

In [17], it is created a new dataset based on data from 2016 to 2018 used routinely collected maternity and birth data for singleton pregnancies that ended in birth at Monash Health, Australia's

largest public health service, at Universal Health Melbourne. Within the framework of the health care system, three maternity hospitals served different ethnic populations.

In another interesting study, a homicidal diabetes dataset was created in [18]. This dataset included a total of 48,502 singleton pregnancies from January 2016 to June 2021 across the Monash Health maternity network. The incidence of GDM was 21.3%. A randomly selected 80% dataset was used for model development and 20% for validation. Performance, including calibration and discrimination performance, was evaluated.

In [19], it is created a dataset obtained from patients attending the Department of Obstetrics and Fetal Medicine at the Hospital Parroquial de San Bernardo, Santiago, Chile [19]. The dataset included data from 1,611 different pregnant patients from 2019 to 2022. The dataset is divided into three parts: the training set (70%), the validation set (10%), and the testing set (20%). In this study, twelve different ML models and their hyperparameters were optimized to achieve early and high predictive performance of GDM. To improve the forecast results, the method of data augmentation was used in training. Three methods were used to select the most suitable variables for GDM prediction. After training, the models with the highest area under the receiver operating characteristic curve were evaluated on the validation set. The models with the best results were evaluated as a measure of generalization performance on the test set.

In [20], data from pregnant Mexican women included in the “Cuido mi Embarazo” (CME) group were used for development (107 cases, 469 controls), and data from the "Monica Pretelini Sáenz" Maternal Perinatal Group was used for the investigation (32 cases, 199 controls) [20]. A 2-hour oral glucose tolerance test (OGTT) with 75 g of glucose at 24-28 weeks of gestation was used to diagnose GDM. A total of 114 single nucleotide polymorphisms with predictive power were selected for evaluation. Blood samples collected during the OGTT were used for SNP analysis. The CME group was randomly divided into a training dataset (70% of the group) and a testing dataset (30% of the group). The training dataset is divided into 10 groups: 9 for building the predictive model and 1 for validation.

In [21], it is created a dataset obtained from a perinatal database for women who gave birth at seven hospitals in four regions of South Korea, under the authority of the Catholic University of Korea from January 2009 to December 2020. Data on mothers’ demographic characteristics, body mass index, blood pressure measurements, blood and urine laboratory tests, diagnoses recorded by doctors, and prescribed medications were collected from the hospital database through electronic medical cards. In this study, a machine learning algorithm was developed to predict gestational diabetes mellitus (GDM) using retrospective data from 34,387 multicenter pregnancies in South Korea.

Figure 2 shows the geographical locations of the countries where the organizations of the authors of the studies analyzed in this review paper.

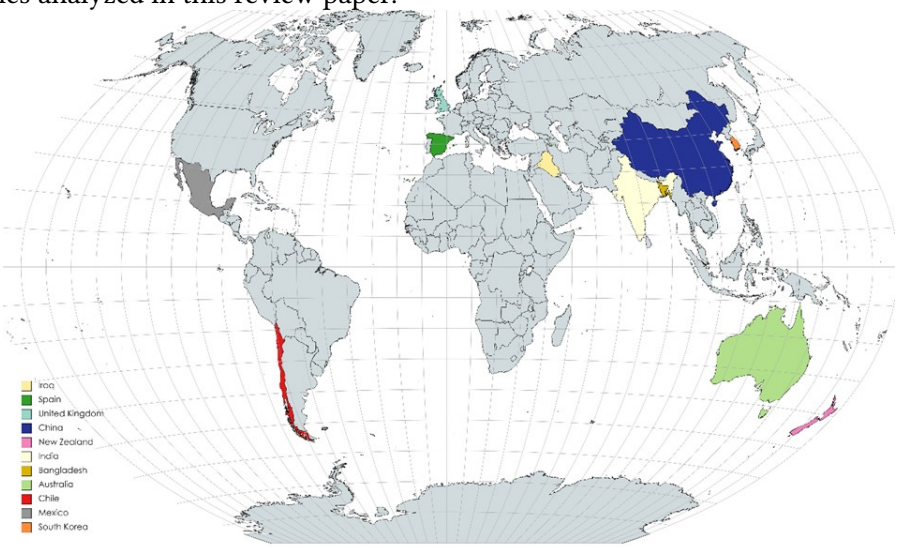

Figure 2: Scope of research to create a dataset for gestational diabetes worldwide (based on studies analyzed in this review paper)

3. Results obtained in the analyzed studies

This section provides information on the results obtained in the articles analyzed in Section 2. The table below shows information about the countries of the organizations of the authors in which the GDM datasets were created, and the number of participants registered in the dataset.

Table 1
Analysis of datasets created for gestational diabetes mellitus

Reference	Country	Hospital	No. of participants in dataset
[7]	Iraq	Kurdistan region laboratories	1012
[8]	Spain	Osakidetza (Basque Health Service)	149 015
[9]	United Kingdom	St. Mary's hospital	26 063
[10]	India	St. Isabella Hospital	600
[11]	China	Tianjin regional laboratories	19 331
[12]	New Zealand	three regional health boards	6 822
[13]	China	West China Second Hospital	10 105
[14]	India	local hospitals of Mysuru	1 352
[15]	Bangladesh	Kurmitola General Hospital	181
[16]	China	one hospital in Beijing	120 396
[17]	Australia	Monash Health	2 880
[18]	Australia	Monash Health maternity hospitals	48 502
[19]	Chile	Hospital Parroquial de San Bernardo	1 611
[20]	Mexico	Maternal Perinatal Hospital Mónica Pretelini	807
[21]	South Korea	seven hospitals in four regions of South Korea	34 387

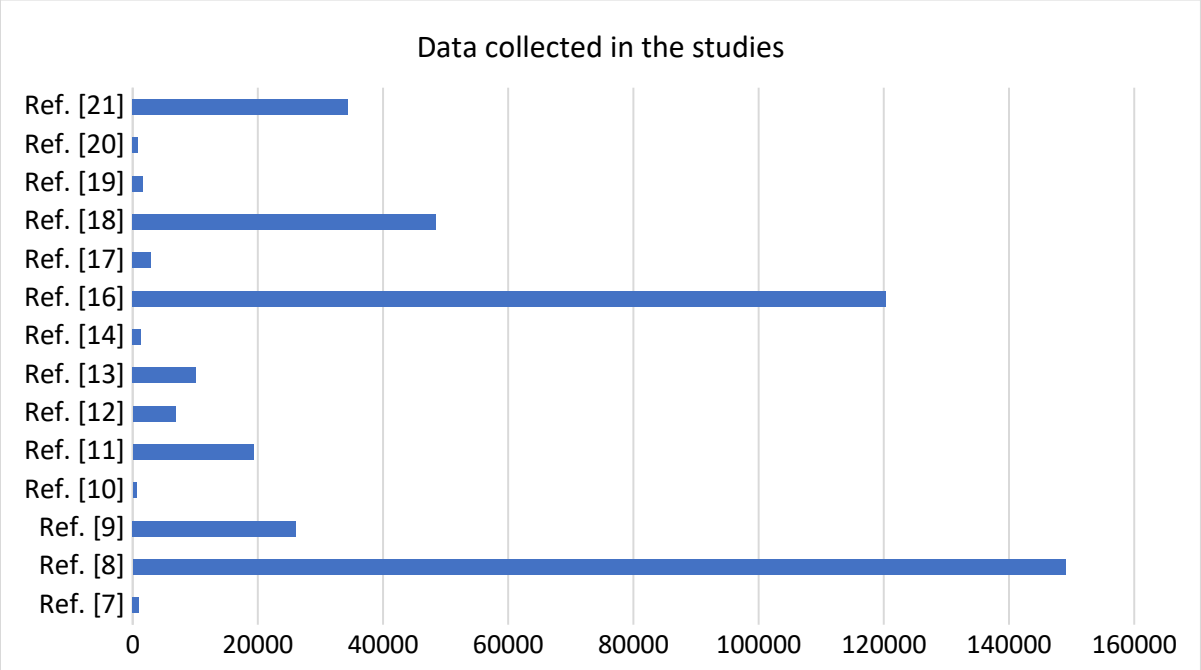

Figure 3: Data collected from the female patients in the studies

Figure 3 depicts the datasets developed in the analyzed studies. If we compare the number of participants in the developed datasets with each other, the five with the highest quantities belong to Ref. [8], Ref. [16], Ref. [18], Ref. [21], and Ref. [9], respectively. Ref. [15], Ref. [10], Ref. [20], Ref. [7], and Ref. [14], were the five lowest quantities on the number of participants in the developed datasets, according to the analyzed studies in this review paper.

Table 2
List of studies evaluating the prediction of gestational diabetes mellitus by machine learning models

Ref.	Algorithm (s)/model (s)	No. of participants in dataset	Results achieved (highest)
[7]	DT, RF, SVM, KNN, LR, and NB	1012	Accuracy: 86.74 %
[8]	Quantum Machine Learning	149 015	Precision: 69 % Accuracy: 69 %
[10]	ID3, Naïve Bayes, C4.5 and Random Tree	600	Accuracy: 93.8 %
[11]	LR and XGBoost	19 331	Specificity: 76.9 % Accuracy: 75.7 %
[13]	LR, Bayesian network, Neural Networks, SVM and CHAID trees	10 105	Accuracy: 90 %
[14]	J48 Decision Tree, RF and NB	1 352	Accuracy: 93 %
[15]	KNN, DT, RF, and NB	181	Accuracy: 81.2 % Precision: 80 % AUC: 84 %
[16]	LR, XGBoost and LightGBM	120 396	Accuracy: 91.7 % Precision: 75.3 % AUC: 92.1 %
[18]	LR, KNN, Gaussian Naïve Bayes (GNB), SVM, DT, multi-layer perceptron (MLP), RF, Extreme randomized tree, AdaBoost, Gradient Boosting, CatBoost, and XGBoost	48 502	Accuracy: 85 % AUC: 93 % Precision: 90 % Recall: 78 % Specificity: 90 %
[19]	MLP and SVM	1 611	Accuracy: 75 % Specificity: 74 % AUC: 81 %
[21]	XGBoost and LightGBM	34 387	AUC: 80.4 %

It can be seen from Table 2 that different types of Machine Learning models/algorithms were used in the analyzed articles. In [7], six ML algorithms were used. In this study, six ML algorithms were used. In this study, the number of participants in the dataset was 1012, and the result was 86.74 % accuracy. In [8], Quantum Machine Learning algorithm was used to develop a prediction model. In this research, the number of participants in the dataset was 149015, and the results were 69 % for both precision and accuracy. In [10], ID3, Naïve Bayes, C4.5 and Random Tree algorithms were used for developing a prediction model. In this work, the number of participants in the dataset was 600, and the result was 93.8 % accuracy. In [11], two ML algorithms, LR and XGBoost, were used to develop a prediction model. In this study, the number of participants in the dataset was 19331, and the results were 76.9 % specificity and 75.7 % accuracy. In [13], five ML algorithms were used for developing a prediction model. In this work, the number of participants in the dataset was 10105, and the result 90 % accuracy. In [14], three ML algorithms were used to develop a prediction model. In this study, the number of participants in the dataset was 1352, and the result was 93 % accuracy. In [15], the prediction model was developed using four ML algorithms: KNN, DT, RF, and NB. In this research, the number of participants in the dataset was 181, and the results were 80 % precision, 84 % AUC and 81.2 % accuracy. A prediction model was developed in [16] using LR, XGBoost, and LightGBM algorithms. In this study, the dataset had 120396 participants. The results showed 91.7% accuracy, 75.3% precision, and 92.1% AUC. In [18], twelve ML algorithms were used to develop a prediction model. In this study, the dataset consisted of 48502 participants. The results showed 85% accuracy, 93% AUC, 90% precision, 78% recall, and 90% specificity. In [19], the prediction model was developed using MLP and SVM algorithms. In this study, the dataset had 1611 participants. The results showed 75 % accuracy, 74 % specificity, and 81% AUC. In [21], XGBoost and LightGBM algorithms were used to develop a

prediction model. In this study, the dataset consisted of 34387 participants, and the resulting AUC was 80.4%.

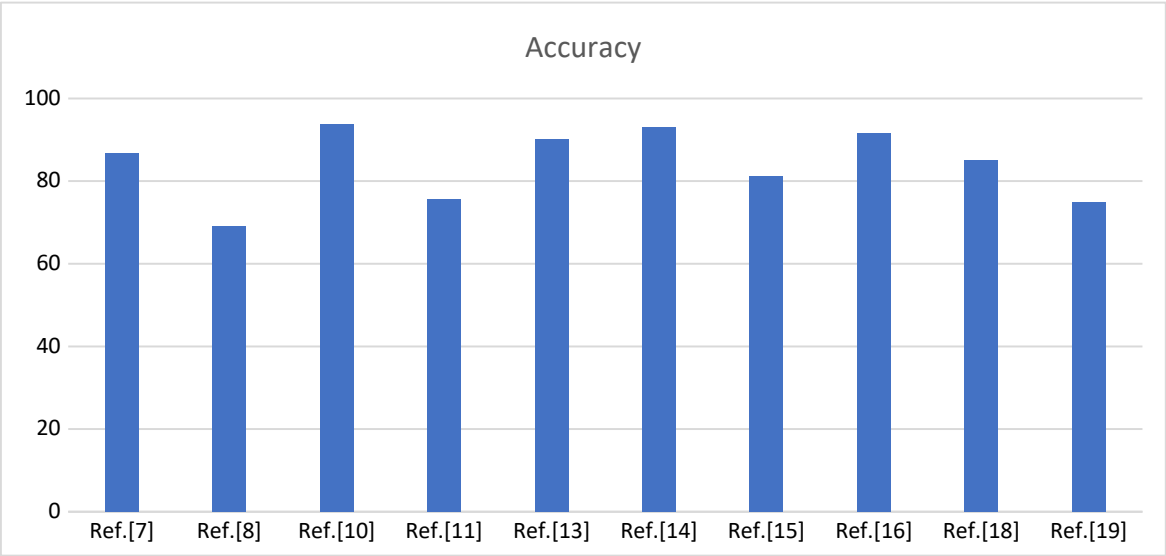

Figure 4: Comparison of analyzed studies based on Accuracy evaluation metric

It can be seen from Figure 4 that in almost all of the analyzed articles, results were obtained based on an accuracy evaluation metric. If we compare the results on accuracy with each other, Ref. [10] had the highest result and Ref. [8] had the lowest result. The results of Ref. [11] and Ref. [19] are close to each other. Also, the results of Ref. [10] and Ref. [14] are close to each other.

4. Conclusion

The paper analyzes the studies conducted on dataset development to assess the risk of developing gestational diabetes mellitus. The results of the study suggest that creating a dataset for assessing the risk of gestational diabetes is a global research topic. The paper also highlights that active research is being conducted on all continents of the world to create a dataset for gestational diabetes mellitus.

Based on the results of the analysis, the following will be conclusions:

Having a large amount of data in a dataset always does not necessarily lead to increased accuracy in machine-learning models. It is important to note that although having more data can be beneficial, it is not the only factor that contributes to model accuracy;

In addition to the dataset, choosing the right prediction models is an important factor in improving model accuracy;

Large amounts of irrelevant or noisy data can mislead the model. Data cleaning and feature engineering are crucial for the effective utilization of large datasets.

In our future work, we plan to create a dataset of women in Uzbekistan with gestational diabetes, using the expertise gained from studying leading scientists' remarkable results during the preparation of this review paper.

References

[1] 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, doi: 10.1016/j.diabres.2019.107843.

[2] K. Nosirov, S. Begmatov, and M. Arabboev, "Design of a model for multi-parametric health monitoring system," *International Conference on Information Science and Communications Technologies, ICISCT 2020.* pp. 1–5, 2020. doi: 10.1109/ICISCT50599.2020.9351522.

[3] P. Bose, S. K. Bandyopadhyay, A. Bhaumik, and S. Poddar, "Female Diabetic Prediction in India Using Different Learning Algorithms," *Univers. J. Public Heal.*, vol. 9, no. 6, pp. 460–471, 2021,

doi: 10.13189/ujph.2021.090614.
[4] M. Arabboev, S. Begmatov, M. Rikhsivoev, S. Saydiakbarov, J. Uraimov, and K. Nosirov, "Gestational diabetes mellitus risk assessment using artificial intelligence: a review," in *International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities*, 2023.
[5] J. Shen *et al.*, "An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus (GDM-AI): Development study," *J. Med. Internet Res.*, vol. 22, no. 9, Sep. 2020, doi: 10.2196/21573.
[6] Z. Zhang *et al.*, "Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis," *J. Med. Internet Res.*, vol. 24, no. 3, 2022, doi: 10.2196/26634.
[7] R. Jader and S. Aminifar, "Predictive Model for Diagnosis of Gestational Diabetes in the Kurdistan Region by a Combination of Clustering and Classification Algorithms: An Ensemble Approach," *Appl. Comput. Intell. Soft Comput.*, vol. 2022, 2022, doi: 10.1155/2022/9749579.
[8] D. Maheshwari, B. Garcia-Zapirain, and D. Sierra-Soso, "Machine learning applied to diabetes dataset using Quantum versus Classical computation," *2020 IEEE Int. Symp. Signal Process. Inf. Technol. ISSPIT 2020*, 2020, doi: 10.1109/ISSPIT51521.2020.9408944.
[9] S. Jeyaparam and R. Agha-Jaffar, "GDM Dataset," *Figshare*, 2023. https://doi.org/10.6084/m9.figshare.21806472.v1
[10] S. Nagarajan, P. Ramasubramanian, and R. . Chandrasekaran, "Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes," *Int. J. Curr. Res. Acad. Rev.*, vol. 2, no. 10, pp. 91–98, 2014.
[11] H. Liu *et al.*, "Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China," *Diabetes. Metab. Res. Rev.*, vol. 37, no. 5, 2021, doi: 10.1002/dmrr.3397.
[12] R. L. Lawrence, C. R. Wall, and F. H. Bloomfield, "Prevalence of gestational diabetes according to commonly used data sources: An observational study," *BMC Pregnancy Childbirth*, vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12884-019-2521-2.
[13] H. Qiu *et al.*, "Electronic Health Record Driven Prediction for Gestational Diabetes Mellitus in Early Pregnancy," *Sci. Rep.*, vol. 7, no. 1, pp. 1–13, 2017, doi: 10.1038/s41598-017-16665-y.
[14] N. Prema and M. Pushpalatha, "Analysis of Risk Factors of Gestational Diabetes Mellitus (GDM) Using Data Mining," *J. Women's Heal. Issues Care*, vol. 8, no. 2, pp. 1–4, 2018, doi: 10.4172/2325-9795.1000329.
[15] B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh," *Inf.*, vol. 11, no. 8, 2020, doi: 10.3390/INFO11080374.
[16] X. Lu, J. Wang, J. Cai, Z. Xing, and J. Huang, "Prediction of Gestational Diabetes and Hypertension Based on Pregnancy Examination Data," *J. Mech. Med. Biol.*, vol. 22, no. 3, pp. 1–18, 2022, doi: 10.1142/S0219519422400012.
[17] S. D. Cooray *et al.*, "Temporal validation and updating of a prediction model for the diagnosis of gestational diabetes mellitus," *J. Clin. Epidemiol.*, vol. 164, pp. 54–64, 2023, doi: 10.1016/j.jclinepi.2023.08.020.
[18] Y. Belsti *et al.*, "Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model," *Int. J. Med. Inform.*, vol. 179, no. September, p. 105228, 2023, doi: 10.1016/j.ijmedinf.2023.105228.
[19] G. Cubillos *et al.*, "Development of machine learning models to predict gestational diabetes risk in the first half of pregnancy," *BMC Pregnancy Childbirth*, vol. 23, no. 1, pp. 1–19, 2023, doi: 10.1186/s12884-023-05766-4.
[20] M. Zulueta *et al.*, "Development and validation of a multivariable genotype-informed gestational diabetes prediction algorithm for clinical use in the Mexican population: Insights into susceptibility mechanisms," *BMJ Open Diabetes Res. Care*, vol. 11, no. 2, 2023, doi: 10.1136/bmjdrc-2022-003046.
[21] B. S. Kang *et al.*, "Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms," *Sci. Rep.*, vol. 13, no. 1, pp. 1–10, 2023, doi: 10.1038/s41598-023-39680-8.