IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Comprehensive Review of Machine learning based Gestational Diabetes Mellitus: The Role of SMOTE and ADSYN

[1] Mrs. Sundari B, [2]Dr.T. NirmalRaj

[1]Research Scholar, [2]Assistant Professor,
[1]Computer Science and Applications,[2] Computer Science and Applications
[1]SCSVMV University, Enathur, Kanchipuram,[2]SCSVMV University, Enathur, Kanchipuram

Abstract: Gestational diabetes mellitus is a form of glucose intolerance that is first recognized during pregnancy and poses significant short and long term risks to both the mother and the fetus. Over the past several decades, diagnostic criteria for GDM have evolved in response to advancements in glucose testing and a growing body of evidence linking maternal hyperglycaemia to adverse perinatal outcomes. It happens when the hormones from the placenta block your ability to use or make insulin. Insulin helps your body maintain the right amount of glucose in your blood. Too much of glucose in your blood can lead to pregnancy complications. Gestational diabetes comes from hormonal changes and the way your body converts foods into energy. Early identification and intervention are essential to reduce adverse outcomes such as preeclampsia, cesarean delivery, macrosomia, and future development of type 2 diabetes mellitus. However, the prediction of GDM is often challenged by class imbalance, high- dimensional clinical data and the complexity of risk factors. The potential of hybrid machine learning approaches for reliable and early GDM prediction, supporting clinical decision making and personalized prenatal care. This review provides a comprehensive overview of class imbalance in GDM datasets and examines two widely adopted oversampling techniques, the Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN), for addressing this challenge. The review discusses the underlying concepts, mechanisms, advantages, and limitations of both techniques, highlighting their ability to generate synthetic minority-class samples and improve data distribution. It also compares the characteristics of SMOTE and ADASYN, emphasizing their applicability to healthcare datasets and their role in enhancing the representation of underrepresented clinical cases. Furthermore, the review summarizes recent research trends, identifies existing challenges associated with synthetic data generation, and outlines future directions for developing more robust and clinically relevant data balancing strategies. By providing a focused analysis of SMOTE and ADASYN in the context of GDM, this review aims to support future research on balanced healthcare datasets and contribute to more reliable and equitable predictive modelling in maternal healthcare.

Keywords: Gestational Diabetes Mellitus, Class Imbalance, SMOTE, ADASYN, Synthetic Oversampling, Healthcare Data Balancing.

I. INTRODUCTION

Gestational Diabetes Mellitus (GDM) is a disorder caused by glucose intolerance that occurs in a woman during pregnancy or in the course of a pregnancy for the first time. This disease is one of the most common complications during pregnancy; however, it causes a number of unfavorable outcomes for both the patient and the fetus, such as preeclampsia, cesarean delivery, macrosomia, neonatal hypoglycemia, as well as a higher risk of getting type 2 diabetes mellitus for both mother and the baby. With the increase of the global obesity rate, the age of patients who are giving birth, and a sedentary way of life, the incidence of this disease has considerably grown.

During recent years, the diagnosis of this disorder has undergone a lot of changes because of the development of new techniques of glucose measurement as well as the knowledge of the connection between maternal hyperglycemia and the outcomes of pregnancy. Various organizations have suggested several criteria for diagnosing this disease which resulted in some differences in the approaches to treatment. However, the common goal of all these approaches is early detection of this disorder. Gestational diabetes will have great negative impact on the physical well-being of both mother and infant. It might increase the risks of adverse events of pregnancy in women in the short-term course, and metabolic disorders and cardiovascular problems in the long-term course in gestational diabetic women and their infants [1–4]. In recent years, with the progress of society and changes in people's lifestyles and eating habits, the proportion of pregnant obese women and older pregnant women keeps increasing year after year [5]. Furthermore, GDM has an increasing incidence in pregnant women [6,7].

As indicated by the diabetes map (the 9th edition) provided by the International Diabetes Federation (IDF) in 2019, 223 million women are affected by diabetes, and the number is predicted to reach 343 million in 2045, with 1/6 pregnancies having GDM [8]. Thus, in light of the increasing threat to public health involving a huge number of patients, it is important to find a solution to the problem [9-11]. Nowadays, routine tests for GDM among pregnant women are done at 24-28 weeks of pregnancy, with the use of the oral 75g glucose tolerance test (OGTT) in the third trimester of pregnancy. Nonetheless, recent researches have indicated that high blood sugar levels in pregnant women can adversely affect the baby even before diagnosis of GDM [12,13]. Early pregnancy intervention in this regard has proved to reduce GDM incidence, thus significantly lowering the rate of complications and improving prognosis [14,15]. Thus, early detection and intervention of GDM play a significant role.

II Literature Review

[1] Karlo abnoosian et al Diabetes mellitus is a life-threatening chronic disease with an increasing global prevalence, requiring early diagnosis and effective treatment to prevent severe health complications. This study proposes an innovative pipeline-based multi-classification framework for diabetes prediction, categorizing patients into three classes: diabetic, non-diabetic, and pre-diabetic, using an imbalanced Iraqi patient diabetes dataset. The proposed framework integrates comprehensive data pre-processing techniques, including duplicate sample removal, attribute transformation, missing value imputation, data normalization and standardization, feature selection, and k-fold cross-validation to enhance model reliability and performance. Several machine learning algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), AdaBoost, and Gaussian Naïve Bayes (GNB), are employed for classification. Furthermore, a weighted ensemble strategy based on the Area Under the Curve (AUC) metric is introduced to effectively address dataset imbalance and improve predictive performance. Experimental results demonstrate that the proposed framework achieves superior performance, with average accuracy, precision, recall, F1-score, and AUC values of 0.9887, 0.9861, 0.9792, 0.9851, and 0.999, respectively.

[2] Somayeh Kianian Big deli, Marjan Ghazisaedi et al analyses to create an early prediction model for GDM n the first trimester of pregnancy. This descriptive study was conducted at the fertility health centre of vali-e-Asr hospital in Tehran, Iran. The dataset was collected from the hospital system from 2020 to 2022. The acquired information underwent pre-processing and six machine learning were created evaluated for GDM prediction in the first trimester pregnancy: Decision Tree, K-Nearest neighbor, Naïve Bayes, Random forest and XG Boost. Models were tested using Accuracy, Precision, Sensitivity and the Area under the Curve. Based on the GTT results the Random model achieved the best performance in GDM prediction with 89% accuracy, 86% precision ,92% recall ,94% AUC using Medical history and clinical findings. The study

highlights the effectiveness of ML approaches with the Random Forest model showing promising performance for the early prediction of GDM during the first trimester.

[3] Valentina Ivanovo and Md Abu Jafar Sujan examines the current landscape of AI and ML applications in GDM prediction, with particular emphasis on models developed using preconception and pregnancy-related data. Various predictive approaches, including Logistic Regression, Support Vector Machines, Decision Trees, Random Forests, Gradient Boosting algorithms, and Neural Networks, are reviewed and compared in terms of predictive performance, interpretability, and clinical applicability. The review highlights the growing adoption of advanced learning algorithms and the increasing availability of electronic health records and multi-source healthcare datasets that support predictive modelling. The findings indicate that ensemble and deep learning methods often demonstrate superior predictive accuracy compared to traditional statistical approaches. However, challenges such as data imbalance, limited external validation, lack of model explain ability, heterogeneous datasets, and inconsistent reporting standards continue to hinder clinical implementation. Furthermore, the integration of longitudinal patient data, explainable AI frameworks, and causal inference techniques remains limited despite their potential to improve model reliability and trustworthiness. This review emphasizes the need for robust validation strategies, transparent reporting practices, and interdisciplinary collaboration among clinicians, data scientists, and healthcare organizations. Future research should focus on developing interpretable, generalizable, and clinically deployable AI-driven prediction models that facilitate personalized risk assessment and early intervention for Gestational Diabetes Mellitus.

[4] Hesham zaky et evaluates gestational diabetes mellitus is one of the most common medical complications during pregnancy. In this paper we propose a ML based techniques for early detection of GDM. For this purpose, we considered Clinical measurements taken during the first trimester to predict the onset of GDM in second trimester. Thus there is a need for the early detection off GDM to avoid critical health conditions in new-borns and post pregnancy complexities of mother. The early detection of GDM We confirmed bio marks that is history of high glucose level/diabetes, insulin and cholesterol which again with the previous studies. The proposed ensemble based model achieved high accuracy in predicting onset of GDM with around 89% accuracy using only the first Trimester data.

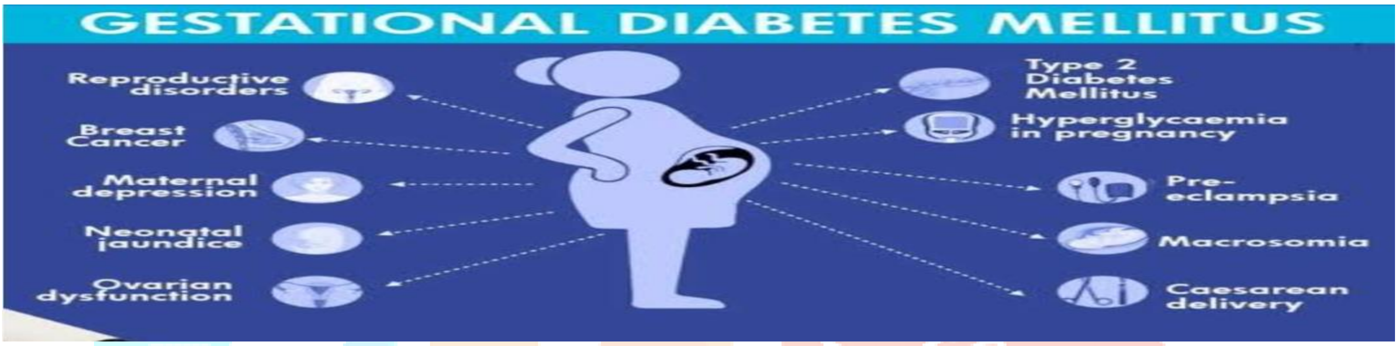
[5] Fang Zhou et al examines the Machine learning techniques have emerged as powerful tools for the early identification and management of Gestational Diabetes Mellitus (GDM). Early prediction can help lower the occurrence of GDM, minimize complications for both mother and baby, and improve pregnancy outcomes. In addition, the influence of clinical factors collected during different stages of pregnancy on model performance is not yet fully understood. In this research, data from 2309 pregnant women collected from two public hospitals in Shenzhen were analysed. Multiple machine learning algorithms were evaluated to develop a reliable stepwise prediction framework consisting of Level A, Level B, and Level C models. The study also included interpretability analysis to better understand the contribution of individual clinical features at different pregnancy stages. Among all the algorithms tested, XGBoost achieved the highest predictive performance. The selected models for Levels A, B, and C obtained accuracy values of 0.922, 0.859, and 0.850, respectively, while the corresponding AUC values were 0.974, 0.924, and 0.913 on the test dataset. These results indicate that the proposed framework can effectively distinguish between GDM and non-GDM cases. To improve model transparency, the researchers applied tree-based feature importance analysis and SHAP techniques. These methods successfully identified well-known GDM risk factors and also revealed varying patterns of feature influence across different stages of pregnancy. The findings provide deeper insight into how clinical parameters contribute to GDM risk over time. This work contributes to improved risk profiling and supports the future integration of machine learning models into healthcare applications.

[6] Kokori et al examines the machine learning (ML) techniques have attracted considerable attention because of their ability to analyse complex clinical datasets and identify predictive patterns that may not be detected using conventional statistical methods. This literature review synthesizes research published between 2000 and September 2023 on the application of ML algorithms for GDM prediction. The reviewed studies demonstrate that ML-based models have substantial potential for identifying women at high risk of developing GDM, particularly during the early stages of pregnancy, thereby enabling timely clinical intervention. The literature also indicates that prediction performance varies across populations, highlighting the importance of developing population-specific models that account for demographic, genetic, and clinical differences rather than relying on universal prediction approaches. Furthermore, many studies report that incorporating diverse

clinical variables, including maternal characteristics, medical history, laboratory findings, and pregnancy-related factors, enhances predictive accuracy and supports more individualized patient management. Although a wide range of ML algorithms has been investigated, challenges remain regarding feature selection, data quality, class imbalance, model interpretability, and external validation. Overall, the existing literature demonstrates the growing role of machine learning in improving GDM prediction while emphasizing the need for robust, clinically applicable, and generalizable predictive models that can facilitate early diagnosis and improve maternal and neonatal health outcomes.

III.Clinical Features of Gestational Diabetes Mellitus (GDM)

Gestational Diabetes Mellitus (GDM) is a metabolic disorder characterized by glucose intolerance that is first detected during pregnancy. The condition usually develops during the second or third trimester as a result of increased insulin resistance induced by placental hormones. Although many women with GDM may not exhibit obvious symptoms, uncontrolled hyperglycaemia can lead to adverse maternal, fetal, and neonatal outcomes.



1.Maternal Clinical Manifestations

1.1 Frequent Urination (Polyuria)

Elevated blood glucose levels exceed the renal threshold, resulting in glucose excretion through urine. This osmotic effect increases urinary output and causes excessive urination.

Clinical implications:

- Urination occurring more frequently than expected during pregnancy
- Increased susceptibility to dehydration

1.2 Excessive Thirst (Polydipsia)

The fluid loss associated with polyuria stimulates the body's thirst mechanism, leading to an increased desire for water intake.

Clinical implications:

- Persistent thirst despite adequate fluid consumption
- Possible indicator of poor glycaemic regulation

1.3 Increased Appetite (Polyphagia)

Due to insulin resistance, glucose uptake by body cells becomes impaired. Consequently, cells experience an energy deficit, which triggers excessive hunger.

Clinical implications:

- Increased food intake
- Potential for excessive gestational weight gain

1.4 Fatigue and Reduced Energy

Inadequate cellular utilization of glucose limits energy production, contributing to persistent tiredness and weakness.

Clinical implications:

- Continuous fatigue
- Reduced ability to perform routine activities

1.5 Visual Disturbances

Hyperglycemia can alter fluid balance within the eye lens, causing temporary changes in vision.

Clinical implications:

- Blurred or fluctuating vision
- Difficulty focusing on objects

1.6 Recurrent Infections

Elevated glucose concentrations support microbial growth and may impair immune responses.

Common infections include:

- Urinary tract infections
- Vulvovaginal candidiasis
- Skin infections

Clinical implications:

- Increased healthcare visits
- Greater risk of pregnancy-related complications

2. Maternal Complications Associated with GDM

2.1 Hypertensive Disorders

Women diagnosed with GDM are at increased risk of developing gestational hypertension and preeclampsia. Chronic hyperglycaemia may contribute to vascular dysfunction and systemic inflammation, leading to elevated blood pressure.

Clinical manifestations:

- Hypertension
- Proteinuria
- Peripheral edema
- Headache and visual symptoms

2.2 Polyhydramnios

Polyhydramnios refers to an excessive volume of amniotic fluid during pregnancy. Maternal hyperglycemia results in fetal hyperglycemia, which stimulates increased fetal urine production and consequently elevates amniotic fluid volume.

Clinical manifestations:

- Rapid increase in abdominal size
- Maternal discomfort
- Breathing difficulties
- Increased likelihood of preterm labor

2.3 Higher Rate of Cesarean Delivery

The presence of fetal overgrowth and obstetric complications frequently increases the need for cesarean section.

Contributing factors:

- Fetal macrosomia
- Prolonged labor
- Shoulder dystocia

3. Fetal Manifestations and Complications

3.1 Fetal Macrosomia

Macrosomia is generally defined as a birth weight exceeding 4,000 g. Excess maternal glucose crosses the placenta and stimulates fetal insulin secretion. Since insulin acts as a growth-promoting hormone, excessive fetal growth may occur.

Potential consequences:

- Difficult vaginal delivery
- Birth trauma
- Increased cesarean section rates

3.2 Shoulder Dystocia

Shoulder dystocia occurs when the fetal shoulders become lodged during delivery, creating an obstetric emergency.

Associated risks:

- Brachial plexus injury
- Bone fractures
- Neonatal complications

3.3 Preterm Birth

Poor glycaemic control has been linked to an increased likelihood of premature delivery.

Consequences include:

- Low birth weight
- Respiratory complications
- Need for neonatal intensive care

3.4 Fetal Distress

Abnormal maternal glucose levels may adversely affect placental function and fetal oxygenation.

Potential outcomes:

- Compromised fetal well-being
- Requirement for emergency intervention

4. Neonatal Manifestations

4.1 Neonatal Hypoglycemia

Neonatal hypoglycemia is among the most frequently observed complications in infants born to mothers with GDM.

Mechanism:
Following birth, the maternal glucose supply ceases abruptly, while elevated fetal insulin levels persist, resulting in reduced neonatal blood glucose concentrations.

Clinical manifestations:

- Tremors or jitteriness
- Poor feeding
- Irritability
- Seizures in severe cases

4.2 Respiratory Distress Syndrome

Fetal hyperinsulinemia may delay pulmonary maturation, increasing the risk of respiratory difficulties after birth.

Clinical manifestations:

- Rapid breathing
- Grunting respirations
- Cyanosis

4.3 Neonatal Jaundice

Infants born to mothers with GDM have an increased tendency to develop hyperbilirubinemia.

Clinical manifestation:

- Yellow discoloration of the skin and sclera

4.4 Metabolic Abnormalities

Additional neonatal complications may include hypocalcemia and polycythemia.

Features include:

- Reduced serum calcium levels
- Increased red blood cell concentration

5. Long-Term Consequences

Maternal Outcomes

Women with a history of GDM have a substantially greater risk of:

- Developing Type 2 Diabetes Mellitus later in life
- Experiencing GDM in subsequent pregnancies
- Developing cardiovascular disorders

Offspring Outcomes

Children born to mothers affected by GDM may face an increased risk of:

- Childhood obesity
- Insulin resistance
- Type 2 Diabetes Mellitus during adolescence or adulthood

IV.Class imbalance problems

In class imbalance (CI), for classification problems, the number of instances of one class is less compared to the other class. Ensemble learning refers to the technique whereby several models are combined in order to build a strong model. It has been used together with data augmentation in solving class imbalance problem. The class imbalance (CI) is one of the difficult issues that occur in machine learning with an imbalanced distribution of instances. The CI issue is the common issue in classification problems (Bi & Zhang, 2018).

Problems Arising Due to Imbalance in Classes

Class imbalance brings about various problems that may greatly influence the effectiveness of machine learning classification algorithms. Some of these problems are discussed below.

1. Bias for the Majority Class

Generally, machine learning algorithms are constructed in such a way that they aim at minimizing classification errors. In an imbalanced dataset, the algorithm is likely to learn from the majority class since there are many instances of data in it. The classifier is therefore biased towards the majority class, while the minority class, which is important in some applications, is overlooked.

2. Poor Minority-Class Classification

Due to the small size of minority class examples, it becomes hard for the classifier to capture their features adequately. Consequently, minority class examples get misclassified as majority class, resulting in many false negatives, a low sensitivity/recall rate, and a poor F1 score. In the context of medicine applications, for instance, missing out on identifying a patient with the disease can have dire implications.

3. Inadequate Performance Measures

The overall accuracy performance metric does not make sense to use in datasets that have class imbalance since it can show good performance even though the classification model is unable to detect instances from the minority class. For instance, in a case where the data has 95% instances from the majority class, a classifier which classifies all the instances as the majority class will score 95% even though it is unable to detect the minority class.

4. Overfitting and Under Fitting Problems

Solutions applied to solve the class imbalance problem might bring new problems. Random oversampling repeats examples of the minority class, which can lead to overfitting because of memorizing these examples. Random under sampling excludes examples of the majority class, thus losing valuable data and leading to under fitting. Despite the sophisticated methods like SMOTE that create synthetic examples of the minority class in order to avoid overfitting, they might cause noise in the data set.

5. Distortion of Decision Boundary

Imbalanced classes may lead to the distortion of the decision boundary whereby the classifier learns to favour one class over the other. As the classifier learns using more data from the majority class than the minority class, it gives more decision space to the majority class and less to the minority class. This will make the classifier perform poorly in identifying minority-class objects.

4.1 SMOTE-Based Data Balancing Techniques

Class imbalance is a common challenge in machine learning, particularly in medical diagnosis, fraud detection, and anomaly detection, where one class contains significantly fewer samples than the other. Traditional machine learning algorithms often become biased toward the majority class, leading to poor identification of minority instances. To overcome this limitation, the **Synthetic Minority Over-Sampling Technique (SMOTE)** has become one of the most widely adopted oversampling methods.

SMOTE generates synthetic minority class samples instead of simply duplicating existing observations. The technique identifies the *k* nearest minority-class neighbors for each minority sample and creates new synthetic instances by interpolating between the selected sample and one of its nearest neighbors. This process expands the minority class within the feature space, improving class balance while reducing the risk of overfitting associated with random oversampling. Since its introduction, SMOTE has been successfully applied in numerous domains, including healthcare, finance, remote sensing, cybersecurity, and engineering. Several enhanced variants of SMOTE have also been proposed to address its limitations in handling noisy data, overlapping classes, and complex decision boundaries.

4.1.1 SMOTE-ENN

SMOTE-ENN is a hybrid sampling technique that combines the oversampling capability of SMOTE with the data cleaning ability of the **Edited Nearest Neighbor (ENN)** algorithm. Initially, SMOTE generates synthetic minority samples to balance the dataset. Subsequently, ENN removes noisy and misclassified observations by comparing each sample with its nearest neighbors. If the class label of a sample differs from the majority of its neighbouring instances, the sample is removed from the dataset. This combined strategy not only balances class distribution but also improves data quality by eliminating ambiguous observations near class boundaries. Consequently, SMOTE-ENN often enhances classifier generalization and has demonstrated strong performance in medical diagnosis, disease prediction, and other noisy classification problems.

4.1.2 K-Means SMOTE

K-Means SMOTE integrates clustering with synthetic oversampling to generate more representative minority samples. The algorithm first partitions the dataset into clusters using the K-Means clustering algorithm. Clusters containing a high proportion of minority samples are then selected for oversampling. Instead of generating synthetic instances uniformly, the algorithm allocates more synthetic samples to sparse minority regions where additional data are most needed. Finally, standard SMOTE is applied within each selected cluster to generate synthetic observations. This localized oversampling strategy preserves the underlying data distribution and reduces the likelihood of producing noisy or overlapping samples. K-Means SMOTE has demonstrated superior performance in medical diagnosis, land-cover classification, cybersecurity, and other highly imbalanced datasets.

4.1.3 SMOTE-SVM

SMOTE-SVM combines SMOTE with the **Support Vector Machine (SVM)** classifier to improve classification performance on imbalanced datasets. SVM is a supervised learning algorithm that constructs an optimal hyperplane to maximize the separation between different classes. Although SVM performs well on balanced datasets, its predictive accuracy decreases when minority class samples are underrepresented. In the SMOTE-SVM approach, SMOTE first generates synthetic minority instances to balance the dataset before training the SVM classifier. The improved class distribution enables SVM to identify a more accurate decision boundary, thereby increasing sensitivity toward minority class instances. This hybrid technique has been successfully applied in medical diagnosis, handwriting recognition, bioinformatics, and spatial data analysis.

4.1.4 Borderline-SMOTE

Borderline-SMOTE is an extension of the traditional SMOTE algorithm that focuses on minority samples located near the decision boundary between classes. These borderline instances are more likely to be misclassified because they are surrounded by majority-class samples. Instead of oversampling all minority instances equally, Borderline-SMOTE identifies minority samples located in "danger" regions and generates synthetic samples specifically around these critical observations. Strengthening the decision boundary improves classifier discrimination and reduces misclassification errors. Borderline-SMOTE has been widely used in medical diagnosis, financial risk assessment, landslide and debris-flow prediction, protein sequence classification, and bankruptcy prediction, where accurate classification of difficult minority cases is particularly important.

Comparison of SMOTE Variants

Technique	Main Idea	Advantages	Limitations
SMOTE	Generates synthetic minority samples using k-nearest neighbors	Simple, reduces class imbalance	May generate overlapping or noisy samples
SMOTE-ENN	Combines SMOTE with ENN data cleaning	Removes noise and improves class separation	Higher computational cost
K-Means SMOTE	Performs clustering before oversampling	Preserves local data structure and reduces noise	Performance depends on appropriate selection of cluster number
SMOTE-SVM	Uses SMOTE before SVM classification	Improves SVM performance on imbalanced data	Sensitive to SVM parameter tuning
Borderline-SMOTE	Oversamples minority samples near class boundaries	Enhances classification of difficult samples	May still be affected by noisy boundary samples

5.2 ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning)

ADASYN is an extension of the Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE creates synthetic observations using all minority class examples as templates. ADASYN, instead, focuses on generating synthetic samples in areas where the minority class is hardest to learn, that is, where minority class is sparsely represented. Like this, ADASYN should enhance the classifier’s ability to better discern the boundaries between the minority class and the majority class examples, or at least, so the theory goes. The ADASYN algorithm relies on k-nearest neighbors (k-NN) to identify regions in the feature space where the minority class is underrepresented.

Below is a step-by-step outline of how ADASYN works:

i) Compute the Class Imbalance

ADASYN first calculates the degree of imbalance in the dataset. This is the ratio between the number of majority and minority class samples. The following image illustrates class imbalance, where yellow dots represent the majority class, while green and purple dots represent the minority class.

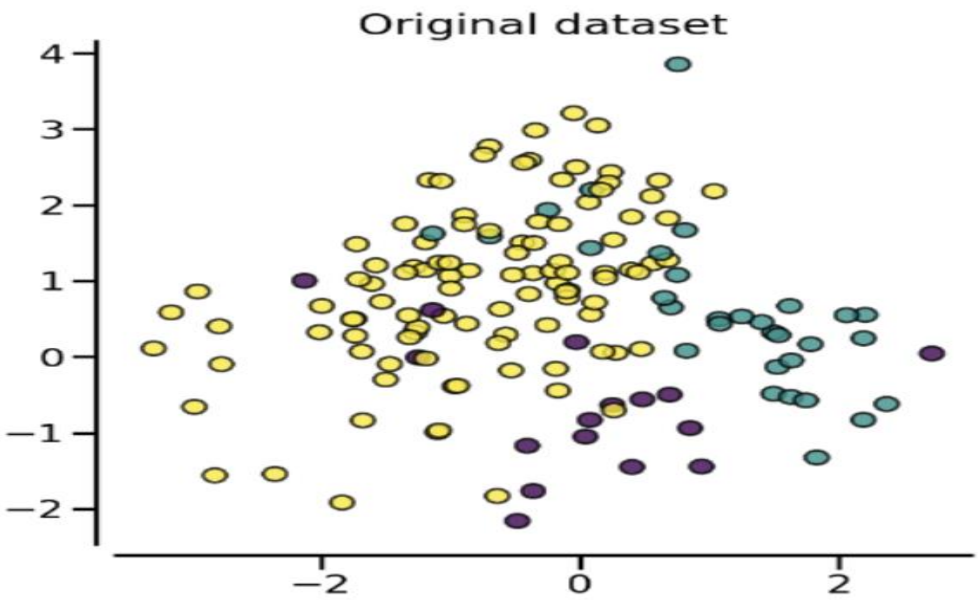


Fig 1: Imbalance dataset class distribution

ii) Identify Difficult-to-Learn Instances

For each minority class sample, ADASYN finds the number of nearest neighbors that belong to the majority class. A higher number of closest neighbors from the majority class suggests that the instance is harder to classify correctly.

The dots circled in red are perfect examples of difficult to learn instances, because they are surrounded by a higher number of majority class instances.

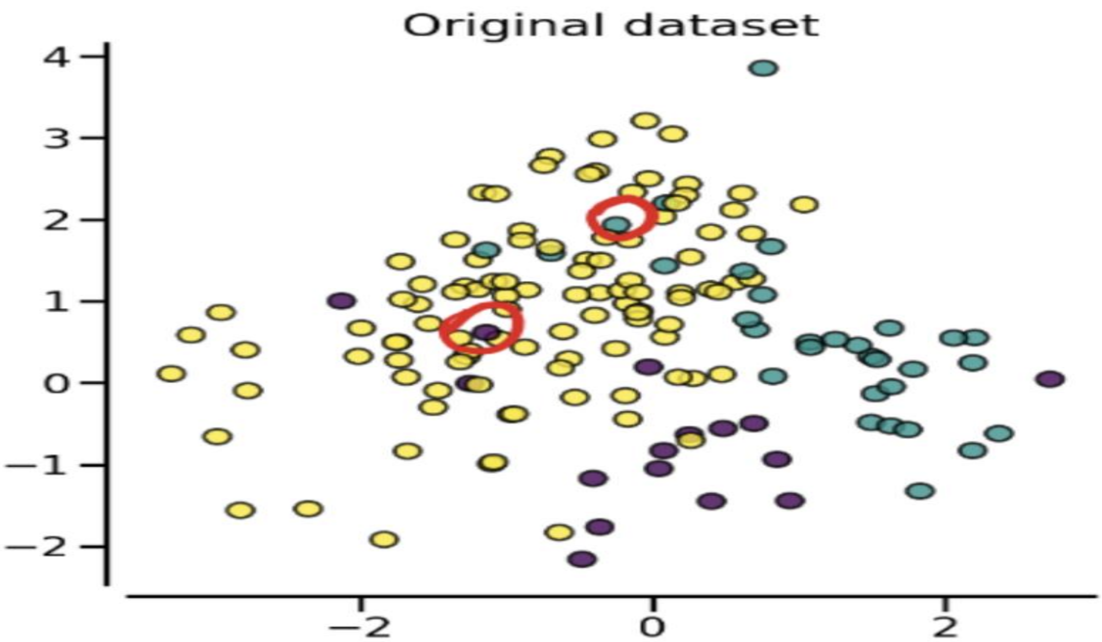


Fig 2- Difficult instances with majority class instances

iii) Compute Sampling Distribution

ADASYN calculates the weight for each minority class instance based on the ratio of majority class neighbors in its local neighbourhood. Specifically, for each minority sample *xi*, the weight *wi* is computed as the number of majority class samples among its k-nearest neighbors, divided by k (the total number of neighbors considered).

This ratio reflects the local imbalance around *xi*. The weights are then normalized so that they sum to 1, forming a probability distribution. Harder-to-learn instances (those with more majority class neighbors) receive higher weights, ensuring that more synthetic samples are generated in their vicinity.

In mathematical terms:

- Let *ri* be the number of majority class samples among the k-nearest neighbors of *xi*.
- The weight *wi* is calculated as wi=ri /k.
- The weights are normalized as wi′=wi / ∑wj, where j varies from 1 to the total number of minority class samples.

This normalized weight wi′ determines the proportion of synthetic samples to be generated for each minority instance.

iv) Balance the Dataset

For each minority class instance *xi*, ADASYN generates synthetic samples based on its normalized weight *wi′*. The number of synthetic samples to be created for *xi* is proportional to *wi′*. To create a new synthetic sample:

- Randomly select one of the k-nearest neighbors of *xi*, denoted as *xzi*.
- Compute the difference vector between *xzi* and *xi*: diff=xzi−xi.
- Multiply this difference vector by a random number λ between 0 and 1: .
- Add this scaled difference vector to *xi* to create the new synthetic sample: *xnew=xi+lambda diff*

This process ensures that synthetic samples are generated along the line between *xi* and its neighbors, focusing more on regions where minority class instances are sparse or harder to classify. The result is an adaptive oversampling technique that improves class balance by emphasizing difficult-to-learn areas.

VI. Conclusion

This review highlights the importance of addressing class imbalance in gestational diabetes mellitus datasets through synthetic oversampling techniques. SMOTE and ADASYN have become widely recognized methods for improving minority class representation, each offering distinct advantages depending on the distribution and complexity of the data. Nevertheless, limitations associated with synthetic data generation indicate that further advancements are needed. Future research should explore intelligent oversampling frameworks, hybrid data balancing methods, privacy-preserving synthetic data generation, and standardized evaluation protocols for healthcare datasets. Establishing robust and clinically validated balancing strategies will strengthen future GDM research and facilitate the development of reliable, equitable, and generalizable healthcare data resources.

VII. References

[1] J. Bi and C. Zhang, "An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme," Knowledge-Based Systems, vol. 158, pp. 81–93, 2018, doi: 10.1016/j.knosys.2018.05.037.

[2] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," Journal of Artificial Intelligence Research, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.

[3] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," Information Sciences, vol. 465, pp. 1–20, 2018.

[4] G. Douzas and F. Bacao, "Geometric SMOTE: A geometrically enhanced drop-in replacement for SMOTE," Information Sciences, vol. 501, pp. 118–135, 2019.

[5] A. Bria, C. Marrocco, and F. Tortorella, “Addressing class imbalance in deep learning for small lesion detection on medical images,” Computers in Biology and Medicine, vol. 120, Art. no. 103735, 2020, doi: 10.1016/j.compbiomed.2020.103735.

[6] J. Li, Y. Chen, C. Tan, Y. Li, F. Gu, Y. Zhang, et al., “Application of the Borderline-SMOTE method in susceptibility assessments of debris flows in Pinggu District, Beijing, China,” Natural Hazards, vol. 105, pp. 2499–2522, 2021.

[7] M. Mishra, B. Patnaik, R. C. Bansal, R. Naidoo, B. Naik, and J. Nayak, “DTCDWT-SMOTE-XGBoost-based islanding detection for distributed generation systems: An approach of class-imbalanced issue,” IEEE Systems Journal, 2021.

[8] Z. Zhang, L. Yang, W. Han, Y. Wu, L. Zhang, C. Gao, K. Jiang, Y. Liu, and H. Wu, “Machine learning prediction models for gestational diabetes mellitus: Meta-analysis,” JMIR Medical Informatics, vol. 10, no. 3, Art. no. e26634, 2022.

[9] C. Chamseddine, N. Mansouri, M. Soui, and M. Abed, “Handling class imbalance in COVID-19 chest X-ray images classification: Using SMOTE and weighted loss,” Applied Soft Computing, vol. 129, Art. no. 109588, 2022.

[10] Q. Ning, X. Zhao, and Z. Ma, “A novel method for identification of glutarylation sites combining Borderline-SMOTE with Tomek links technique in imbalanced data,” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 19, no. 5, pp. 2632–2641, 2022.

[11] P. Puri and M. K. Gupta, “Improved hybrid bag-boost ensemble with K-means-SMOTE–ENN technique for handling noisy class imbalanced data,” The Computer Journal, vol. 65, no. 1, pp. 124–138, 2022.

[12] A. Sharma, P. K. Singh, and R. Chandra, “SMOTified-GAN for class imbalanced pattern classification problems,” IEEE Access, 2022.

[13] F. Zhou, X. Ran, F. Song, Q. Wu, Y. Jia, Y. Liang, S. Chen, G. Zhang, J. Dong, and Y. Wang, “A stepwise prediction and interpretation of gestational diabetes mellitus: Foster the practical application of machine learning in clinical decision,” Heliyon, vol. 10, no. 10, Art. no. e31350, 2024.

[14] E. Kokori, G. Olatunji, N. Aderinto, I. Muogbo, I. J. Ogieuhi, D. Isarinade, B. Ukoaka, A. Akinmeji, I. Ajayi, E. Chidiogo, O. Samuel, H. Nurudeen-Busari, A. O. Muili, and D. B. Olawade, “The role of machine learning algorithms in detection of gestational diabetes: A narrative review of current evidence,” Clinical Diabetes and Endocrinology, vol. 10, p. 18, 2024, doi: 10.1186/s40842-024-00176-7.

[15] M. Zhao, Z. Yao, Y. Zhang, L. Ma, W. Pang, S. Ma, Y. Xu, and L. Wei, “The progression of gestational diabetes mellitus to type 2 diabetes: A systematic review and meta-analysis,” Diabetology & Metabolic Syndrome, vol. 16, Art. no. 133, 2024.

[16] S. K. Bigdeli, M. Ghazisaedi, S. M. Ayyoubzadeh, S. Hantoushzadeh, and M. Ahmadi, “Predicting gestational diabetes mellitus in the first trimester using machine learning algorithms: A cross-sectional study at a hospital fertility health center in Iran,” BMC Medical Informatics and Decision Making, vol. 25, no. 1, Art. no. 3, 2025, doi: 10.1186/s12911-024-02799-3.

[17] H. Zaky, E. Fthenou, L. Srour, T. Farrell, M. Bashir, N. El Hajj, and T. Alam, “Machine learning based model for the early detection of gestational diabetes mellitus,” BMC Medical Informatics and Decision Making, vol. 25, no. 1, Art. no. 130, 2025, doi: 10.1186/s12911-025-02947-3.

[18] A. Khan, O. Chaudhari, and R. Chandra, “A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation,” *Expert Systems with Applications*, vol. 244, Art. no. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.

[19] R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: Illustration and simulation using logistic regression," J. Am. Med. Inform. Assoc., vol. 29, no. 9, pp. 1525–1534, 2022, doi: 10.1093/jamia/ocac093.

[20] Y. F. Chen et al., "Design of a clinical decision support system for fracture prediction using imbalanced dataset," J. Healthcare Eng., vol. 2018, Art. no. 9621640, 2018.

[21] Y. Zhu and C. Zhang, "Prevalence of gestational diabetes and risk of progression to Type 2 diabetes: A global perspective," Curr. Diabetes Rep., vol. 16, no. 1, Art. no. 7, 2016.

[22] H. J. Holmes, J. Y. Lo, D. D. McIntire, and B. M. Casey, "Prediction of diabetes recurrence in women with class A1 (diet-treated) gestational diabetes," Am. J. Perinatol., vol. 27, no. 1, pp. 47–52, 2010.

[23] M. A. Sargın et al., "Neutrophil-to-lymphocyte and platelet-to-lymphocyte ratios: Are they useful for predicting gestational diabetes mellitus during pregnancy?," Ther. Clin. Risk Manag., vol. 12, pp. 657–665, 2016.

[24] K. K. Ryckman, C. N. Spracklen, C. J. Smith, J. G. Robinson, and A. F. Saftlas, "Maternal lipid levels during pregnancy and gestational diabetes: A systematic review and meta-analysis," BJOG, vol. 122, no. 5, pp. 643–651, 2015.

[25] X. Wang et al., "Exposure to multiple metals in early pregnancy and gestational diabetes mellitus: A prospective cohort study," Environ. Int., vol. 135, Art. no. 105370, 2020.

[26] Y. Tokuda et al., "Assessing items on the SF-8 Japanese version for health-related quality of life: A psychometric analysis based on the nominal categories model of item response theory," Value Health, vol. 12, no. 4, pp. 568–573, 2009.

[27] S. Fukuhara and Y. Suzukamo, Manual of the SF-8 Japanese. Kyoto, Japan: Institute for Health Outcomes & Process Evaluation Research, 2004, pp. 71–116.

[28] D. Marchetti, D. Carrozzino, F. Fraticelli, M. Fulcheri, and E. Vitacolonna, "Quality of life in women with gestational diabetes mellitus: A systematic review," J. Diabetes Res., vol. 2017, Art. no. 7058082, 2017.

[29] L. Zhang, W. Zheng, W. Huang, X. Liang, and G. Li, "Differing risk factors for new onset and recurrent gestational diabetes mellitus in multipara women: A cohort study," BMC Endocr. Disord., vol. 22, Art. no. 3, 2022.

[30] C. N. Hales et al., "Fetal and infant growth and impaired glucose tolerance at age 64," BMJ, vol. 303, pp. 1019–1022, 1991.

[31] A. Mahendra and C. H. D. Fall, "Maternal vitamin D deficiency and GDM risk: Evidence for the case of investing more attention in antenatal clinics," Proc. Nutr. Soc., vol. 20, pp. 1–7, 2021.

[32] X. Chen et al., "Association of maternal folate and vitamin B12 in early pregnancy with gestational diabetes mellitus: A prospective cohort study," Diabetes Care, vol. 44, no. 1, pp. 217–223, 2021.

[33] M. Morikawa et al., "Change in the number of patients after the adoption of IADPSG criteria for hyperglycemia during pregnancy in Japanese women," Diabetes Res. Clin. Pract., vol. 90, no. 3, pp. 339–342, 2010.

[34] T. A. Furukawa et al., "The performance of the Japanese version of the K6 and K10 in the World Mental Health Survey Japan," Int. J. Methods Psychiatr. Res., vol. 17, no. 3, pp. 152–158, 2008.

[35] C. L. Craig et al., "International physical activity questionnaire: 12-country reliability and validity," Med. Sci. Sports Exerc., vol. 35, no. 8, pp. 1381–1395, 2003.

[36] N. Murase, T. Katsumura, C. Ueda, S. Inoue, and T. Shimomitsu, "International standardization of physical activity level: Reliability and validity study of the Japanese version of the International Physical Activity Questionnaire (IPAQ)," J. Health Welf. Stat. (Kosei no Shihyo), vol. 49, pp. 1–9, 2003 (in Japanese).

[37] M. Iwai-Shimada et al., "Questionnaire results on exposure characteristics of pregnant women participating in the Japan Environment and Children Study (JECS)," Environ. Health Prev. Med., vol. 23, Art. no. 45, 2018.

[38] Y. Yokoyama et al., "Validity of short and long self-administered food frequency questionnaires in ranking dietary intake in middle-aged and elderly Japanese in the Japan Public Health Center-Based Prospective Study for the Next Generation (JPHC-NEXT) Protocol Area," J. Epidemiol., vol. 26, no. 8, pp. 420–432, 2016.

[39] C. Kim, D. K. Berger, and S. Chamany, "Recurrence of gestational diabetes mellitus: A systematic review," Diabetes Care, vol. 30, no. 5, pp. 1314–1319, 2007.

[40] J. S. Cramer, "The origins of logistic regression," Tinbergen Institute Working Paper No. 2002-119/4, 2002. [Online]. Available: https://doi.org/10.2139/ssrn.360300