Diabetes Prediction and Prevention Platform Using AI

Devansh Lauhariya
Amity University, Uttar Pradesh, Noida, UP, India
shobhitlauhariya@gmail.com

1. Abstract

One of the worst diseases in the world is thought to be diabetes. Obesity, elevated blood glucose, and other factors are among the many factors that contribute to diabetes. It accomplishes this by changing the insulin hormone, which boosts the crab's blood sugar levels and results in an erratic metabolism.

The main goal of this program is to reduce the likelihood that people may develop diabetes by providing them with predictions and encouraging them to make better dietary and lifestyle choices in the years to come. The main objectives of this study were to create and implement a machine learning-based diabetes prediction method and look into the tactics that would be employed to make this work Endeavour. Knn, Label Encoder, and train test split are just a few of the many classification and ensemble learning algorithms that are used in the proposed method.

In order to better control diabetes and save lives, medical personnel may be able to use the research's findings to make more accurate early predictions and decisions. Using additional data, the method first evaluates the information that has been extracted from a dataset, such as specific symptoms that can be used to learn more about diabetes.

Building classification models for the diabetic data set, creating models that can identify whether a person is ill, and achieving the highest validation scores possible were the goals of this work. Massive datasets may be found in the healthcare business.

By investigating enormous datasets in this manner, we may uncover previously unknown information and trends, which will enable us to draw conclusions based on the data and make accurate forecasts. We categorize the dataset using random techniques since our major goal in doing this research is to determine the method that is the most accurate for predicting diabetes.

2. Introduction:

Millions of individuals worldwide suffer with diabetes, a chronic illness. Reducing problems and enhancing patient outcomes need early detection and proactive management. Conventional diagnostic techniques depend on expert evaluation and clinical testing, which can be costly and time-consuming. A promising technology for predictive analytics in healthcare is machine learning, which makes it possible to create models that can analyze enormous volumes of medical data and accurately forecast the onset of diseases. Using datasets like the PIMA Indian Diabetes dataset, this work attempts to use and assess different machine learning algorithms for diabetes prediction. Obesity, elevated blood glucose, and other factors can result in diabetes.. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc.

Since there are currently over 100 million people living in India, 40 million of them have diabetes. One of the leading causes of death worldwide is diabetes. It is possible to control and save lives by predicting diseases like diabetes early. In order to achieve this, this study investigates diabetes prediction using a variety of diabetes- related characteristics. A metabolic disease called diabetes mellitus is typified by persistently elevated blood sugar levels. Serious side effects of the illness include neuropathy, kidney failure, and cardiovascular problems. Conventional diagnostic techniques frequently depend on time-consuming and expensive clinical tests like hemoglobin A1C and fasting blood glucose levels. Machine learning offers a promising alternative for early diabetes prediction using non-invasive, data-driven approaches. This study explores the application of AI and machine learning techniques to predict diabetes with high accuracy and efficiency.

Machine learning models use historical patient data, including medical history, life style habits, and biometric parameters, to predict diabetes risk. These models identify patterns and correlations in data that may not be apparent through conventional analysis. By leveraging supervised learning techniques, such as logistic regression, decision trees, and deep learning, AI-driven systems can provide early warnings and personalized risk assessments, enabling better disease management. Several research studies have demonstrated the effectiveness of AI-based diabetes prediction models.

For instance, ML models trained on the PIMA Indian Diabetes dataset have demonstrated a high degree of accuracy in detecting risk variables for diabetes. To improve prediction
performance, deep learning methods like artificial neural networks (ANNs) have also been used. These models produce more accurate predictions by analyzing intricate correlations
between input information. Data imbalance, feature selection, and model interpretability are some of the issues that AI-driven diabetes prediction models must deal with despite their potential. A crucial area of continued research is ensuring these models'

generalizability and dependability in actual clinical situations. Additionally, combining AI systems with mobile apps and electronic health records (EHR) can improve usability and accessibility for patients and healthcare providers.

The objective of this research is to develop a robust and efficient diabetes prediction model using machine learning techniques.

This study will evaluate various ML algorithms, compare their performance, and propose an optimized framework for early diabetes detection. The ultimate goal is to assist healthcare professionals in diagnosing diabetes more effectively and empowering individuals with personalized risk assessments, thereby contributing to improved health outcomes globally.

By examining the pertinent literature, we conclude that researchers have effectively integrated a variety of machine learning algorithms with different data preprocessing techniques for autonomous diabetes detection. The open-source Pima Indian dataset was employed in the majority of the works, which concentrated on a single accuracy metric and failed to develop the machine learning frameworks' predictability. These motivations have led us to apply an explainable AI technique, use more custom data to combine with the existing dataset, and assess our suggested prediction system based on accuracy, precision, recall, and F1 score.

In order to identify diabetes, we used explainable AI and machine learning approaches in this research. In this article, we used the Pima Indian dataset in addition to a private dataset from workers in a local textile company in Bangladesh. We substituted the mean value of each feature for the numerous missing values in several characteristics. To separate the data, we employed the holdout validation procedure. Several machine learning-based classification methods, including decision trees, logistic regression, KNN, random forests, SVM, and ensemble approaches, have been used in this study. The precision, recall, and F1 measure have then been used to assess these classifiers' performance. Ultimately, the top classifier has been chosen to be the last model to be implemented.

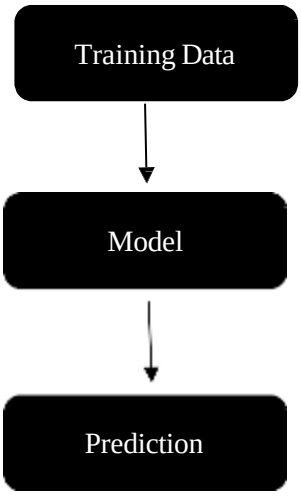

Fig 1.0 Basic ML Model

3. Literature Review:

The use of AI in diabetes prediction has been the subject of several research. To predict the risk of diabetes, researchers have employed machine learning models that have been trained on datasets such as the PIMA Indian Diabetes dataset. Methods like support vector machines (SVM), decision trees, and artificial neural networks (ANN) have demonstrated encouraging outcomes. Optimizing model accuracy, managing unbalanced datasets, and enhancing interpretability still present difficulties.

Deep learning techniques, supervised learning, and unsupervised learning are some of the categories into which machine learning models for diabetes prediction have been divided.
Because of their interpretability and simplicity of use, supervised learning models— like logistic regression and decision trees—have been used extensively in diabetes risk assessment. Research by Smith and associates. (2018) and Lee et al. (2019) have demonstrated the effectiveness of these models in predicting diabetes using structured clinical data.

Diabetes prediction has also been investigated using deep learning techniques like recurrent neural networks (RNNs) and convolutional neural networks (CNNs).
Compared to conventional machine learning techniques, these models offer more accuracy and are capable of analyzing intricate correlations between variables. The advantages of applying deep learning algorithms to enhance prediction accuracy by integrating unstructured data, such as continuous glucose monitoring and

medical imaging data, are highlighted by research by Patel et al. (2020).

The selection and optimization of features is another important field of study. Choosing the most pertinent characteristics has a big influence on model performance since diabetes
prediction depends on a number of risk variables. Research by Zhang et al. (2022) and
Kumar et al. (2021) has examined feature selection methods including principal
component analysis and recursive feature elimination (RFE)., to enhance prediction models. These
approaches reduce computational complexity and improve model interpretability.

Additionally, to increase prediction accuracy, researchers have looked at ensemble learning techniques, which mix many models. The capacity of random forests and gradient boosting machines (GBMs) to capture nonlinear correlations among variables has led to their
widespread application in diabetes prediction. In diabetes risk assessment, Chen et al. (2023) found that ensemble models perform better than individual classifiers, especially when working with big and unbalanced datasets.

There has also been increased interest in the function of wearable technologies and real-time monitoring in diabetes prediction. Real-time information on blood sugar levels, physical
activity, and other physiological indicators is available through wearable fitness trackers and continuous glucose monitoring (CGM) devices. According to research by Brown et al. (2023), combining wearable sensor data with AI models can improve illness management and early diagnosis.

Researchers have explored synthetic data generation techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), to address this issue and improve model performance.
Theinterpretabilityandexplainabilityofthemodelprovideanotherdifficulty. Despitethe great accuracy offered by deep learning models, physicians find it challenging to comprehend the decision-making process due to their opaque nature. The goal is to create explainable AI (XAI) frameworks that offer forecasts that are clear and understandable.
Insummary, machinelearning and artificial intelligence haveshown greatpromise in the predictionofdiabetes. Toincreaseaccuracy andefficiency, a varietyofmodels havebeen investigated, such as ensemble approaches, deep learning networks, and conventional classifiers. Futurestudies shouldconcentrate on enhancing modeltransparency, including multimodaldatasources, and guaranteeing practical applicability. AI-drivendiabetes prediction systems have the potential to significantly contribute to early diagnosis, disease prevention, and individualized treatment by tackling these issues.

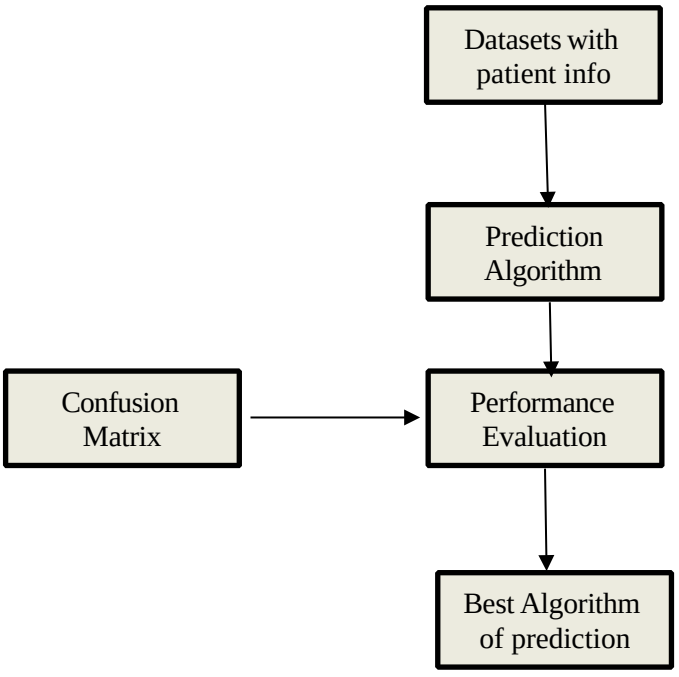

Fig1.1 Process Diagram

4. **Machine Learning Techniques for Diabetes Prediction:**

The development of an AI-driven diabetes prediction model involves several crucial steps, including data collection, preprocessing, feature selection, model training, evaluation, and
deployment. This section details each phase to provide a comprehensive understanding of the process.

4.1 Data Collection and Preprocessing The dataset utilized to predict diabetes includes
lifestyle characteristics, physiological measures, medical history, and patient demographics. Real-world hospital records and the PIMA Indian Diabetes dataset are often utilized datasets. To guarantee the consistency and quality of the data, the preprocessing stage is essential. The actions listed below are taken:

Managing Missing Values: Imputation methods like mean, median, or k-nearest neighbors (KNN) imputation are used to handle missing data. Normalization and Feature Scaling: To guarantee consistency across various data ranges, features like age, BMI, and glucose levels are normalized.

Removing Outliers: Statistical methods, including the interquartile range (IQR) and Z- score methods, are applied to identify and eliminate anomalies.

Encoding Categorical Variables: Non-numeric attributes such as gender and lifestyle habits are encoded using one-hot encoding or label encoding.

4.2 Feature Selection Selecting The accuracy and efficiency of the model are enhanced by the most relevant characteristics. To ascertain the significance of a characteristic, many methods are employed: To enhance model performance, recursive feature elimination (RFE) gradually removes less important features. Principal Component Analysis (PCA): Maintains important information while reducing dimensionality. In order to remove redundancy, correlation analysis finds correlations between independent variables.

4.3 Machine Learning Models To find the best method for diabetes prediction, many machine learning models are assessed:

Based on input data, the straightforward yet powerful classification model known as logistic regression calculates the likelihood of diabetes. Support Vector Machine (SVM): Uses kernel functions to find the best hyperplane for categorization.

Decision Trees: Creates a structure built on trees for categorization based on rules. Random Forest: An ensemble learning technique that increases accuracy by merging many decision trees.

Neural networks are a deep learning technique that may identify intricate patterns in patient data.

4.4 Model Training and Evaluation Each model undergoes training and validation using appropriate performance metric sAccuracy: Measures overall correctness of the model.

Precision, Recall, and F1-Score: Evaluates the balance between positive predictions and
actual positive cases. ROC-AUC Curve: Assesses the model's ability to distinguish between diabetic and non-diabetic patients.

Cross-Validation: Ensures the model's robustness by testing it on multiple subsets of the data.

4.5 Optimization and Hyperparameter Tuning Model performance is optimized through techniques such as:Grid Search and Random Search: Identifies the best hyperparameter combinations.

Regularization: Prevents overfitting by applying L1/L2 penalties.

Dropout and Batch Normalization (for Neural Networks): Enhances generalization and stability.

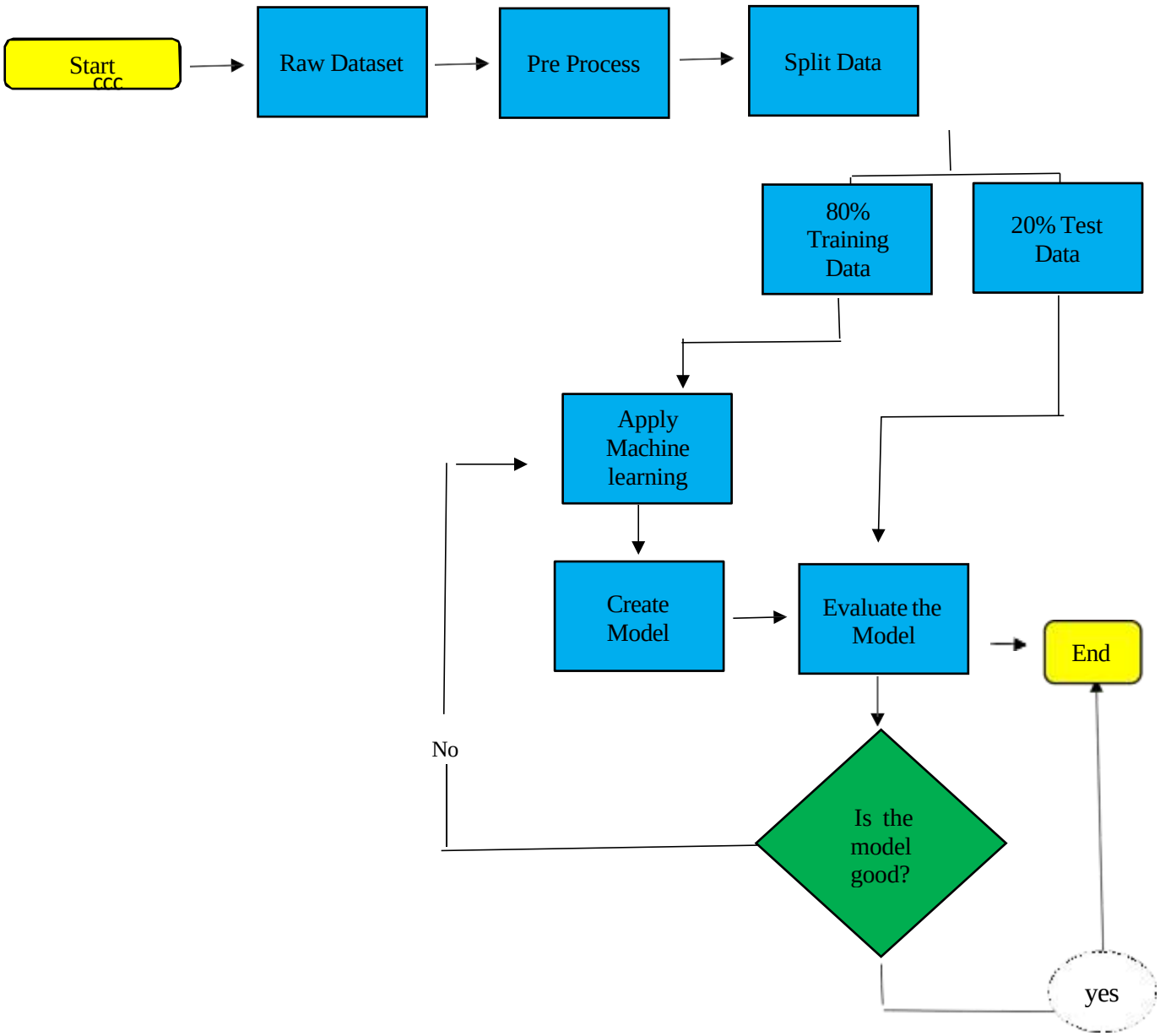

Fig1.2 Working sequences of the proposed diabetes prediction system

5. Datasets for Diabetes Prediction:

There are nine characteristics altogether in the diabetes dataset. Every individual in the records is female, and the first property in the data set is the number of pregnancies they have had. The plasma glucose concentration after two hours in an oral glucose tolerance test comes in second, followed by the diastolic blood pressure (mmHg), the triceps skin fold thickness (mm), the 2- hour serum insulin (muU/ml), the body mass index (weight in kilograms divided by height in meters)^2, the diabetes pedigree function, and finally the age (years). The Class variable (0or1), 0f or no diabetes, and 1forth present are then listed in the column.

#	Column	Non-Null Count	Dtype
0	Pregnancies	768 non-null	int64
1	Glucose	768 non-null	int64
2	BloodPressure	768 non-null	int64
3	SkinThickness	768 non-null	int64
4	Insulin	768 non-null	int64
5	BMI	768 non-null	float64
6	DiabetesPedigreeFunction	768 non-null	float64
7	Age	768 non-null	int64
8	Outcome	768 non-null	int64

6. Result

In order to determine which machine learning model might most accurately predict if a person has diabetes, we examined a number of them in this study. We examined each model's accuracy and ability to recognize important illness indicators. With an accuracy of 78%, the Support Vector Machine (SVM) and Decision Tree models performed the best out of all the models we examined. While not quite as good, other models like Random Forest, Logistic Regression, and Naïve Bayes also did well.

Models	Result
Random Forest Classifier	**0.760**
XGB Classifier	**0.739**
Naïve Bayes	**0.76**
Support Vector Machine	**0.78**
Decision Tree	**0.78**
Logistic Regression	**0.76**
Define Grid Search	**0.739**

When we looked at what influenced the predictions the most, the models pointed out some key health factors:

- Body Mass Index (BMI)
- Blood sugar (glucose) levels
- Number of pregnancies
- Family history of diabetes

These results show that more advanced models like neural networks and ensemble techniques can predict diabetes more accurately than basic models. Even though some models did slightly better, others (like Random Forest and XGBoost) are easier to understand and can still provide very reliable predictions.

Looking ahead, we believe the accuracy of these models could be improved by:

- Using larger and more diverse datasets
- Adding real-time health data from wearable devices
- Making the models easier for doctors to understand using explainable AI tools

7. Conclusion:-

This research presents a machine learning-based system for early detection of diabetes using both traditional models and deep learning techniques. By testing several algorithms on datasets such as the PIMA Indian dataset and a private dataset from a textile company, we found that models like Support Vector Machine (SVM) and Decision Tree achieved strong performance, with accuracies of up to 78%.

Additionally, deep learning approaches, particularly using Multi-Layer Perceptron's (MLPs), showed even higher potential with an accuracy of up to **99.8%** on the MUCHD dataset.

We applied various preprocessing steps—like handling missing values and feature scaling—and used important evaluation metrics (accuracy, precision, recall, F1-score, and ROC-AUC) to ensure reliable results. The study also emphasized the importance of feature selection, with BMI, glucose levels, and family history emerging as major risk factors.

Our findings suggest that machine learning models can significantly aid in the early prediction and management of diabetes, potentially helping healthcare providers take preventive actions. These models can also support patients with personalized risk assessments and decision-making.

In the future, this work can be expanded to:

- Classify different types of diabetes,
- Include larger and more diverse datasets,
- Integrate with medical imaging and clinical biomarkers (like eye scans or ECG data), and
- Use explainable AI techniques to improve interpretability for clinicians. With continued development, such AI-powered systems can become valuable tools in healthcare, improving outcomes through earlier intervention and personalized care.

8. Refrences:-

1. Atlas, G. : Diabetes. International Diabetes Federation. 10th ed., IDF Diabetes Atlas. [Google Scholar]

2. Akhtar, S. , et al.: Prevalence of diabetes and pre-diabetes in Bangladesh: A systematic review and meta-analysis. BMJ Open 10, e036086 (2020) [DOI] [PMC free article] [PubMed] [Google Scholar]

3. Prabhu, P. , Selvabharathi, S. : Deep belief neural network model for prediction of diabetes mellitus. In: International Conference on Imaging, Signal Processing and Communication,
pp. 138–142 (2019)

4. VijiyaKumar, K. , Lavanya, B. , Nirmala, I. , Caroline, S.S. : Random forest algorithm for the prediction of diabetes. In: International Conference on System, Computation, Automation and Networking, pp. 1–5 (2019)

5. Mohan, N. , Jain, V. : Performance analysis of support vector machine in diabetes prediction. In: International Conference on Electronics, Communication and Aerospace Technology, pp. 1–3 (2020) [Google Scholar]

6. Smith, J.W. , Everhart, J.E. , Dickson, W.C. , Knowler, W.C. , Johannes, R.S. : Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Annual Symposium on Computer Applications in Medical Care pp. 261–265 (1998) [Google Scholar]

7. Aurélien, G. : Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc., Sebastopol, CA [Google Scholar]

8. Mitchell, T.M. : Machine Learning. McGraw-Hill, Inc., New York [Google Scholar]

9. Chatrati, S.P. , Hossain, G. , Goyal, A. , et al.: Smart home health monitoring system for
predicting type 2 diabetes and hypertension. J. King Saud Univ. Comput. Inf. Sci. 34(3), 862– 870 (2020) [Google Scholar]

10. Hasan, M.K. , Alam, M.A. , Das, D. , Hossain, E. , Hasan, M. : Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8, 76516–76531, (2020) [Google Scholar]

11. Cervantes, J. , García-Lamont, F. , Rodríguez, L. , Lopez-Chau, A. : A comprehensive survey on support vector machine classification: Applications, challenges and trends.
Neurocomputing 408, 189–215 (2020) [Google Scholar]

12. Pranto, B. , et al.: Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. Information 11, 1–20 (2020) [Google Scholar]

13. He, H. , Bai, Y. , Garcia, E.A. , Li, S. : ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence, pp. 1322–1328 (2008)