iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** X **Month of publication:** October 2023

DOI: https://doi.org/10.22214/ijraset.2023.56214

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue X Oct 2023- Available at www.ijraset.com

An In-depth Comparison of Diabetes Detection Using Machine Learning

Shirin Pinjari[1], Nilesh Vani[2]
[1]*PG Student, GF's Godavari College of Engineering, Jalgaon, India, 425002*
[2]*Associate Professor, GF's Godavari College of Engineering, Jalgaon, India, 425002*

Abstract: ***The timely detection of diabetes is of paramount importance for improving patient outcomes and reducing the overall healthcare burden. This research paper delves into the use of a wide range of machine-learning algorithms to achieve accurate diabetes detection. By leveraging diverse methods including K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression, this study enhances our understanding of how these algorithms perform in the field of diabetes diagnosis. The paper presents comprehensive experimental results and comparative analyses, shedding light on the strengths and limitations of each algorithm within this critical medical domain. The findings contribute to the ongoing effort to develop effective tools for the early detection and management of diabetes, ultimately benefiting both patients and the healthcare system.***
Keywords: data mining, machine learning, Diabetes, Decision Trees, Healthcare, Logistic Regression, Naïve Bayes, Random Forrest, SVM.

I. INTRODUCTION

Diabetes, known as Diabetes Ailment (DA), is a metabolic disorder marked by persistent high levels of blood glucose, accompanied by disruptions in the metabolism of carbohydrates, fats, and proteins. This ailment encompasses three primary types: Type 1 DA, Type 2 DA, and gestational diabetes.

Type 1 DA arises due to the body's inability to generate insulin, necessitating insulin injections or the use of insulin pumps. This variant was formerly termed "insulin-dependent diabetes Ailment" (IDDA) or "juvenile diabetes."

Type 2 DA emerges from insulin resistance, a condition in which cells do not properly respond to insulin, often accompanied by insufficient insulin production. In the past, it was referred to as a non-insulin-dependent diabetes ailment (NIDDA) or "adult-onset diabetes."

The third major form, gestational diabetes, materializes when pregnant individuals without a prior diabetes diagnosis experience elevated blood glucose levels. It can potentially precede the development of Type 2 Diabetes Mellitus (DM). In the year 2000, it was estimated that approximately 171 million individuals worldwide were affected by diabetes, constituting about 2.8% of the global population. Among the various types of diabetes, Type 2 diabetes was the most prevalent on a global scale. Data from 2007 revealed that the five countries with the highest numbers of diagnosed diabetes cases were India (40.9 million), China (38.9 million), the United States (19.2 million), Russia (9.6 million), and Germany (7.4 million) [1].

With the increasing volume of unstructured diabetic data originating from the healthcare industry and various other sources, there is a pressing need to organize and quantify this data effectively. Technological advancements have made it possible to amalgamate robust diabetic data sharing and electronic communication systems, enhancing access to healthcare services for patients at all levels of care. This necessitates the consolidation of all patient data into a single repository. The deployment of a Health Information Exchange (HIE) serves as a solution, as it can collect clinical information from disparate sources and integrate it into a unified patient health record accessible securely by all care providers. Predictive Analysis, a method employing techniques from data mining, statistics, and game theory, leverages historical and current data alongside statistical and analytical models to forecast future events. In the healthcare sector, big data analytics can play a vital role in making significant predictions and informed decisions.

This paper introduces the use of predictive analysis algorithms within a Hadoop/Map Reduce environment to predict prevalent diabetes types, associated complications, and appropriate treatment methods. Through this analysis, the system aims to offer an efficient approach to diagnosing and care for patients, emphasizing factors like affordability and availability to achieve better patient outcomes. Diabetes mellitus stands as a chronic metabolic disorder, which has experienced a remarkable surge in global prevalence within recent decades. The consequences of diabetes going undiagnosed or being inadequately managed are profound, resulting in severe health complications and heightened mortality rates.

The field of machine learning, renowned for its capacity to decipher intricate data patterns, has emerged as a promising tool for timely disease detection and accurate diagnosis. This research zeroes in on the potential of a diverse spectrum of machine learning algorithms, harnessed to precisely pinpoint cases of diabetes. The central objective of this study is to assess the efficacy of diverse algorithms in the realm of diabetes diagnosis. Our investigation delves into an extensive array of algorithms, encompassing K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression. Each algorithm brings forth a distinctive classification approach, and we intend to foster understanding through a comprehensive comparative analysis of their respective advantages and application scopes in the context of diabetes detection.

II. RELATED WORK

In the paper authored by P. Yasodha and M. Kannan [2], the authors employ classification techniques on various types of datasets to determine whether an individual has diabetes or not. The dataset used for diabetic patients is sourced from a hospital warehouse, comprising 249 instances with seven attributes. These instances in the dataset are categorized into two groups, namely blood tests and urine tests. In the work by N. Niyati Gupta, A. Rawal, and V. Narasimhan [3], the primary objective is to assess the accuracy, sensitivity, and specificity percentages of various classification methods. The study also involves comparing and analyzing the results of these classification methods in different software tools, including WEKA, Rapidminer, and Matlab, using the same parameters (accuracy, sensitivity, and specificity). The authors applied JRIP, Jgraft, and BayesNet algorithms for their analysis.
Both papers appear to focus on utilizing classification techniques and machine learning algorithms to address the issue of diabetes diagnosis and evaluation of classification model performance. The first paper uses a specific dataset from a hospital, while the second paper explores the performance of various classification methods across different software platforms.

III. PROPOSED METHODOLOGY

A. Data Collection

Gather a comprehensive dataset that includes medical records of individuals, including features such as age, gender, family history of diabetes, lifestyle factors (diet, exercise), and most importantly, clinical measurements like fasting blood glucose levels, HbA1c levels, BMI, etc.

B. Data Preprocessing

Handle missing data: Impute missing values using techniques like mean imputation or advanced imputation methods.
Normalize or standardize numerical features to bring them to a consistent scale.
Encode categorical variables into numerical values using techniques like one-hot encoding or label encoding.

C. Feature Selection/Engineering

Select relevant features using techniques like feature importance ranking, correlation analysis, or domain knowledge.
Create new features that may be informative, such as BMI categories or a diabetes risk score.

D. Model Selection

Choose appropriate machine learning algorithms for classification, such as logistic regression, decision trees, random forests, support vector machines, or deep learning models.
Experiment with different models to determine which one performs best for your specific dataset is explained below:

1) *K-Nearest Neighbors (KNN):* KNeighbors Classifier can be imported from the following machine-learning library: Sci-kit-learn Select the appropriate value for ‘k’ (number of neighbors): Cross-validate KNN model. Fit KNN model to the training dataset. Train KNN model to predict diabetes outcome on the test dataset. Evaluate KNN model using: Accuracy Precision Reconcordance F1-score
2) *Logistic Regression:* Logistic Regression is a type of machine learning library that you can import from your existing library. Set up the model with the appropriate hyperparameters and fit it to your training data. Then, use it to predict diabetes outcomes on your testing dataset. You can also use it to measure your model's accuracy, precision and recall, as well as to get an F1 score.
3) *Naive Bayes:* The Gaussian Natural Language Model (GaussianNB) from an existing machine learning library using the Naive Bayes machine learning library. After that, the model is created and adjusted to the training set of data. On a test dataset, the model's outcomes can then be utilised to forecast diabetes outcomes. The model can also be assessed for accuracy, precision accuracy, recall accuracy, and accuracy using the F1-score.

4) *Linear Discriminant Analysis (LDA):* Linear Discriminant Analysis can be imported from a machine learning library. Make a new instance of the model for linear discriminant analysis. model fit using the training set of data. On the basis of the test dataset, forecast diabetes outcomes. Use accuracy, precision, recall, and F1-score to assess the model.
5) *Decision Tree:* Import Decision Tree Classifier from a machine learning library. Initialize the decision tree model non the training set of data, fit the model. Using the test dataset, forecast the results of diabetes. Use the model's accuracy, precision, recall, and F1-score to evaluate it.
6) *Random Forest:* a machine learning library's Random Forest Classifier should be imported. Make an instance of the Random Forest model with the proper hyperparameters. Utilise the training dataset to train the model. Using the test dataset, forecast the results of diabetes. Use the model's accuracy, precision, recall, and F1-score to evaluate it.
7) *Ada Boost with Random Forest:* Add the Random Forest Classifier and AdaBoost Classifier libraries. For the basic estimator, create an instance of the Random Forest Classifier. Using the base estimator, instantiate the AdaBoost Classifier. On the training set of data, fit the model. Using the test dataset, forecast the results of diabetes. Use the model's accuracy, precision, recall, and F1-score to evaluate it.
8) *AdaBoost with Logistic Regression:* Import Logistic Regression and AdaBoost Classifier from the library. For the base estimator, create a Logistic Regression object. Using the base estimator, instantiate the AdaBoost Classifier. on the training set of data, fit the model. Using the test dataset, forecast the results of diabetes. Use the model's accuracy, precision, recall, and F1-score to evaluate it.

E. Dataset Description

It's great to have an understanding of your dataset before diving into building machine-learning models. Here's a breakdown of the information you've provided about your dataset:

1) *No Null Values:* The dataset doesn't contain any missing values, which is a positive aspect of modeling since missing data can often pose challenges.
2) *Dimensions of the Dataset:* The dataset consists of 253,680 rows and 22 columns, meaning you have a substantial amount of data to work with. However, having a high number of dimensions compared to the number of samples can sometimes lead to challenges in some machine learning algorithms, so feature selection or dimensionality reduction techniques might be considered.
3) *Class Distribution:* The dataset has two classes: "Diabetes" and "Non-Diabetes." The class distribution is as follows:
- Diabetes Class: 35,346 instances
- Non-Diabetes Class: 218,334 instances
4) *Sampled Subset:* For analysis, a random sample of 5,000 rows from each class, totaling 10,000 rows. This can be a useful approach when working with large datasets, as it reduces computation time while still providing a representative subset for modeling.
5) *Supervised Learning:* The dataset is suitable for supervised learning, where you have labeled data with input features and corresponding target labels (in this case, the "Diabetes" or "Non-Diabetes" class).

IV. RESULT EVALUATION AND DISCUSSION

1) Decision Tree Classifier Accuracy for Different Depths. Max Accuracy is 73.04% for depth 7.

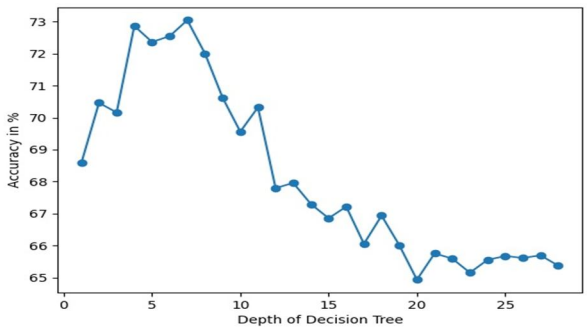


Fig 1: Accuracy of Decision Tree for different depths

2) Adaboost classifier with Base Estimator = Logistic Regression. Max Accuracy is 72.92% for learning rate 1 and number of estimators 11.

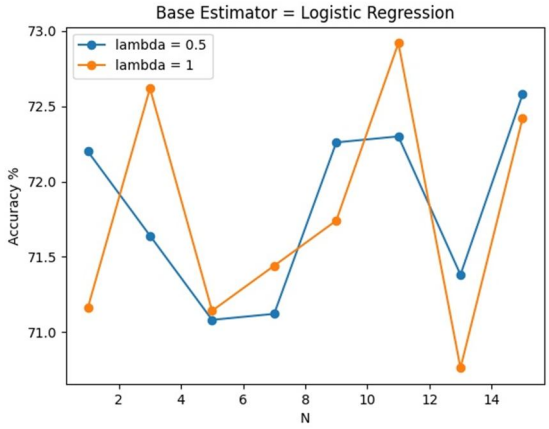


Fig 2: Accuracy of Adaboost classifier with Logistic Regression

3) Adaboost classifier with Base Estimator = Naïve Bayesian. Max Accuracy is 71.90% for learning rate 1 and number of estimators 7.

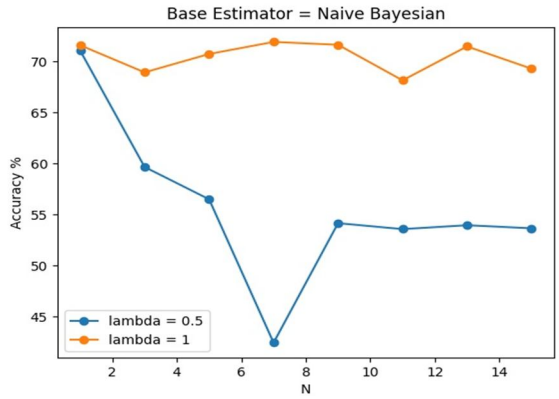


Fig 3: Accuracy of Adaboost classifier with Naïve Bayesian

4) Adaboost classifier with Base Estimator = Random Forest. Max Accuracy is 75.54% for a learning rate of 0.5 and several estimators 13.

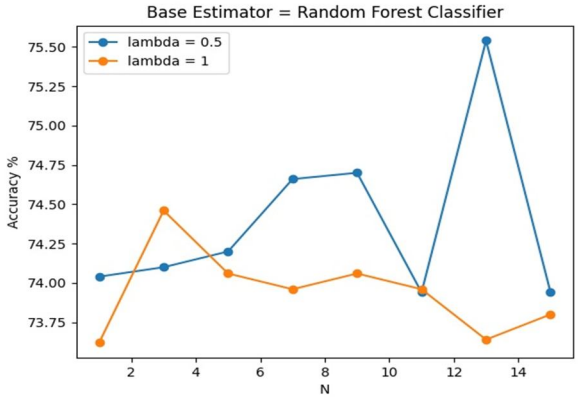


Fig 4: Accuracy of Adaboost classifier with Random Forest

5) K Nearest Neighbour for P = 1.5. The highest accuracy is 71.48% for 11 neighbors.

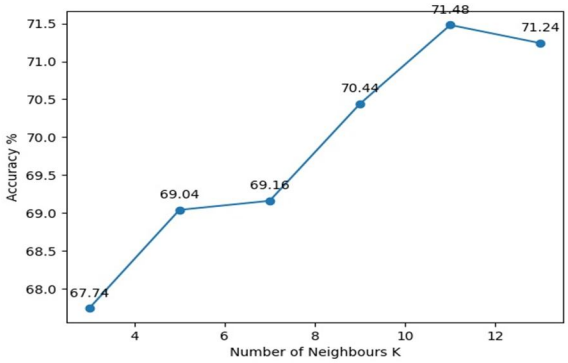


Fig 5: Accuracy of K Nearest Neighbour

6) Accuracy for Random Forest Classifier with different number of estimators and different depths. Max accuracy is 74.58% for 15 estimators and depth: 7.

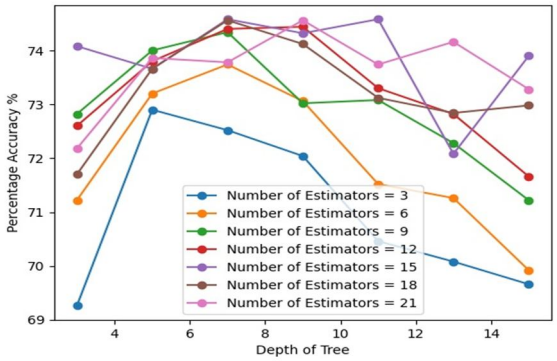


Fig 6: Accuracy of Random Forest Classifier with different number of Estimator

Table 1 Comparative analysis of Machine Learning Classifier concerning Accuracy

Machine Learning Classifier	Accuracy %	Machine Learning Classifier	Accuracy %
KNN with P = 1	71.34%	Adaboost with Naïve Bayesian	71.90%
KNN with P = 1.5	71.48%	Linear SVM	74.84%
KNN with P = 2	71.34%	Gaussian SVM	73.88%
KNN with P = 3	71.22%	Polynomial SVM Degree 2	73.88%
Logistic Regression	74.36%	Polynomial SVM Degree 3	73.88%
Naïve Bayesian	71.90%	Polynomial SVM Degree 4	73.88%
Linear Discriminant	75.00%	Polynomial SVM Degree 5	73.88%
Quadratic Discriminant	70.72%	Random Forest	74.58%
Decision Tree	73.04%	Adaboost with Logistic Regression	72.92%
Adaboost with Random Forest	75.54%		

V. CONCLUSION

In this study, we embarked on an exploration of various machine learning algorithms for the critical task of diabetes detection. Through extensive experimentation, we gained insights into the performance and applicability of K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Tree, Random Forest, AdaBoost with Random Forest, and AdaBoost with Logistic Regression in diagnosing diabetes.

Our findings revealed that each algorithm presents a unique approach to diabetes detection, with varying degrees of accuracy and interpretability. KNN exhibited competitive results by leveraging neighborhood information, while Logistic Regression provided a simpler yet effective model. Naive Bayes demonstrated its strengths in probabilistic modeling, while Linear Discriminant Analysis showcased its potential in capturing class separation. Decision Tree offered transparency in decision-making, and Random Forest excelled in ensemble-based classification. AdaBoost with Random Forest and AdaBoost with Logistic Regression showcased the power of boosting techniques in improving the performance of base classifiers**.**

VI. FUTURE SCOPE

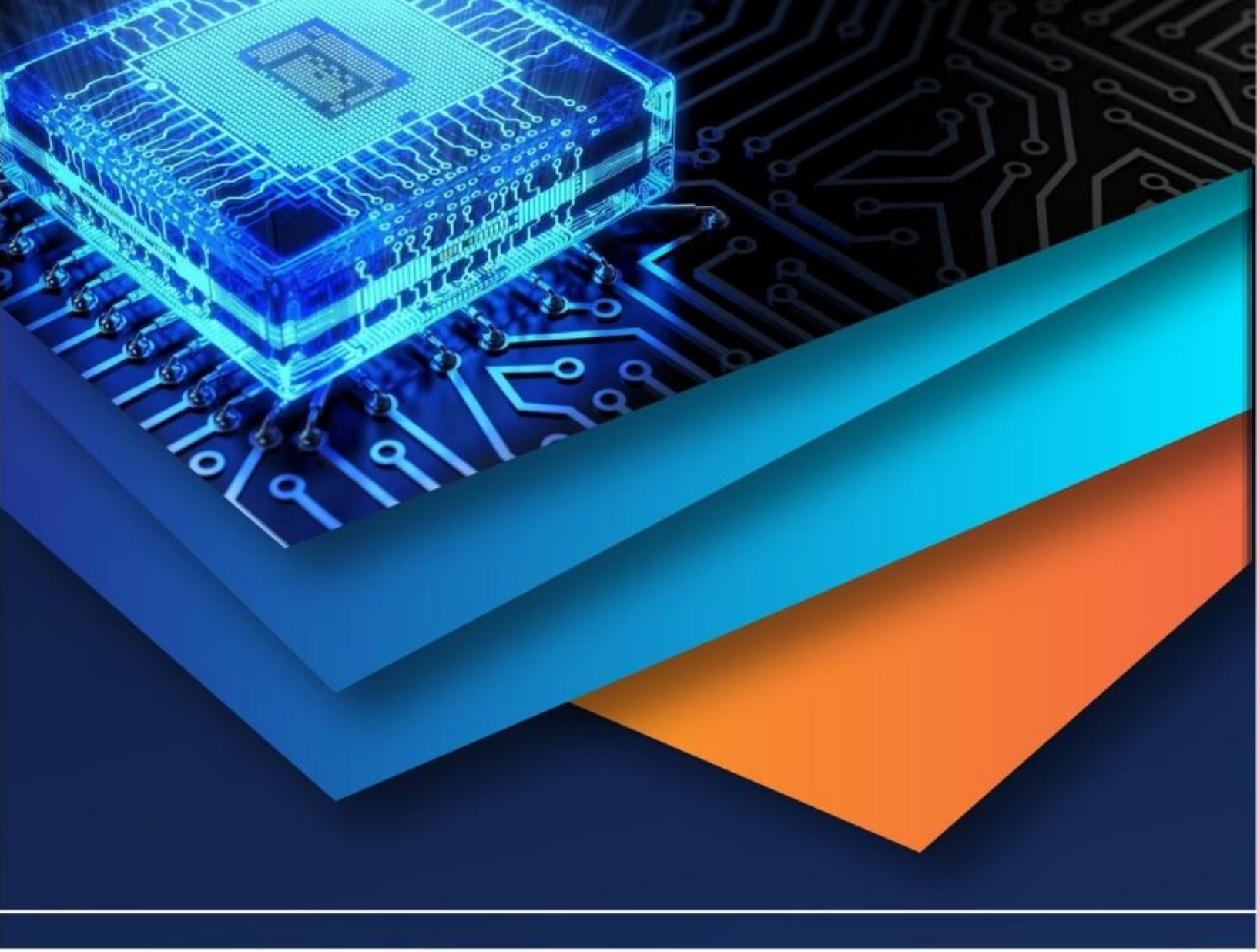
The present study opens avenues for further research and development in the domain of diabetes detection using machine learning. Some potential directions for future investigations include:

1) *Ensemble Refinement:* Investigate more sophisticated ensemble strategies, hybridizing different algorithms to exploit their combined strengths and mitigate weaknesses.
2) *Feature Engineering:* Explore advanced feature engineering techniques to enhance the discriminatory power of the models, potentially using domain-specific knowledge.
3) *Deep Learning Integration:* Integrate deep learning architectures, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to leverage the power of complex feature extraction.
4) *Interpretability Enhancement:* Develop techniques to enhance the interpretability of complex models like Random Forest or AdaBoost, making their decisions more comprehensible to medical professionals.
5) *Domain Adaptation:* Extend the study to adapt the trained models to different demographic groups, as diabetes detection might vary across populations.
6) *Real-time Deployment:* Implement a user-friendly interface for healthcare practitioners to input patient data and receive automated diabetes risk assessments.
7) *Multi-Class Classification:* Extend the binary classification task to multi-class classification, accommodating more comprehensive diagnosis scenarios.
8) *Longitudinal Data Analysis:* Explore the effectiveness of the proposed algorithms in analyzing longitudinal patient data for predicting diabetes progression and identifying risk factors.

REFERENCES

[1] P. T. Katzmarzyk, C. L. Craig, and L. Gauvin, "Adiposity, physical fitness, and incident diabetes: The physical activity longitudinal study," Diabetologia, vol. 50, no. 3, pp. 538–544, Mar. 2007.
[2] Z. Xu, X. Qi, A. K. Dahl, and W. Xu, "Waist-to-height ratio is the best indicator for undiagnosed type 2 diabetes," Diabetic Med., vol. 30, no. 6, pp. e201–e207, Jun. 2013.
[3] R. N. Feng, C. Zhao, C. Wang, Y. C. Niu, K. Li, F. C. Guo, S. T. Li, C. H. Sun, and Y. Li, "BMI is strongly associated with hypertension and waist circumference is strongly associated with type 2 diabetes and dyslipidemia, in northern Chinese adults," J. Epidemiol., vol. 22, no. 4, pp. 317–323, May 2012.
[4] A. Berber, R. G´omez-Santos, G. Fangh¨anel, and L. S´anchez-Reyes, "Anthropometric indexes in the prediction of type 2 diabetes mellitus, hypertension and dyslipidemia in a Mexican population," Int. J. Obes. Relat Metab. Disorders, vol. 25, no. 12, pp. 1794–1799, Dec. 2001.
[5] B. Balkau, D. Sapinho, A. Petrella, L. Mhamdi, M. Cailleau, D. Arondel, and M. A. Charles, D. E. S. I. R. Study Group, "Prescreening tools for diabetes and obesity-associated dyslipidemia: Comparing BMI, waist and waist-hip ratio. The D.E.S.I.R. Study," Eur. J. Clin. Nutr., vol. 60, no. 3, pp. 295–304, Mar. 2006.
[6] I. S. Okosun, K. M. Chandra, S. Choi, J. Christman, G. E. Dever, and T. E. Prewitt, "Hypertension and type 2 diabetes comorbidity in adults in the United States: risk of overall and regional adiposity," Obes. Res., vol. 9, no. 1, pp. 1–9, Jan. 2001.
[7] L. A. Sargeant, F. I. Bennett, T. E. Forrester, R. S. Cooper, and R. J. Wilks, "Predicting incident diabetes in Jamaica: the role of anthropometry," Obes. Res., vol. 10, no. 8, pp. 792–798, Aug. 2002.
[8] N. T. Duc Son le, T. T. Hanh, K. Kusama, D. Kunii, T. Sakai, N. T. Hung, and S. Yamamoto, "Anthropometric characteristics, dietary patterns and risk of type 2 diabetes mellitus in Vietnam," J. Amer. Coll. Nutr., vol. 24, no. 4, pp. 229–234, Aug. 2005.
[9] G. T. Ko, J. C. Chan, C. S. Cockram, and J. Woo, "Prediction of hypertension, diabetes, dyslipidemia or albuminuria using simple anthropometric indexes in Hong Kong Chinese," Int. J. Obes. Relat. Metab. Disorders, vol. 23, no. 11, pp. 1136–1142, Nov. 1999.
[10] M. B. Snijder, P. Z. Zimmet, M. Visser, J. M. Dekker, J. C. Seidell, and J. E. Shaw, "Independent and opposite associations of waist and hip circumferences with diabetes, hypertension and dyslipidemia: The AusDiab study," Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 3, pp. 402–409, Mar. 2004.

[11] B. J. Lee, B. Ku, J. Nam, D. D. Pham, and J. Y. Kim, “Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes,” IEEE J. Biomed. Health Information., vol. 18, no. 2, pp. 555–561, Mar. 2014.
[12] L. de Koning, H. C. Gerstein, J. Bosch, R. Diaz, V. Mohan, G. Dagenais, S. Yusuf, and S. S. Anand, EpiDREAM Investigators, “Anthropometric measures and glucose levels in a large multi-ethnic cohort of individuals at risk of developing type 2 diabetes,” Diabetologia, vol. 53, no. 7, pp. 1322–1330, Jul. 2010.
[13] I. S. Okosuna and J.M.Boltrib, “Abdominal obesity, hypertriglyceridemia, hypertriglyceridemic waist phenotype and risk of type 2 diabetes in American adults,” Diabetes Metab. Syndrome, vol. 2, no. 4, pp. 273–281, Dec. 2008.
[14] Z. Yu, L. Sun, Q. Qi, H. Wu, L. Lu, C. Liu, H. Li, and X. Lin, “Hypertriglyceridemic waist, cytokines and hyperglycemia in Chinese," Eur. J. Clin. Invest., vol. 42, no. 10, pp. 1100–1111, Oct. 2012.
[15] T. Du, X. Sun, R. Huo, and X. Yu, "Visceral adiposity index, hypertriglyceridemic waist and risk of diabetes: The China health and nutrition survey 2009,” Int. J. Obes. (Lond.), vol. 38, no. 6, pp. 840–847, Jun. 2014.
[16] M. Solati, A. Ghanbarian, M. Rahmani, N. Sarbazi, S. Allahverdian, and F. Azizi, "Cardiovascular risk factors in males with hypertriglycemicwaist (Tehran lipid and glucose study),” Int. J. Obes. Relat. Metab. Disorders, vol. 28, no. 5, pp. 706–709, May 2004.
[17] I. Lemieux, A. Pascot, C. Couillard, B. Lamarche, A. Tchernof, N. Alm´eras, J. Bergeron, D. Gaudet, G. Tremblay, D. Prud’homme, A. Nadeau, and J. P. Despr´es, “Hypertriglyceridemic waist: A marker of the atherogenic metabolic triad (hyperinsulinemia; hyper apolipoprotein B; small, dense LDL) in men?" Circulation, vol. 102, no. 2, pp. 179–184, Jul. 2000.
[18] L. B. Tank´o, Y. Z. Bagger, G. Qin, P. Alexandersen, P. J. Larsen, and C. Christiansen, “Enlarged waist combined with elevated triglycerides is a strong predictor of accelerated atherogenesis and related cardiovascular mortality in postmenopausal women,” Circulation, vol. 111, no. 15, pp. 1883–1890, Apr. 2005.
[19] I. F. Gazi, T. D. Filippatos, V. Tsimihodimos, V. G. Saougos, E. N. Liberopoulos, D. P. Mikhailidis, A. D. Tselepis, and M. Elisaf, “The hypertriglyceridemic waist phenotype is a predictor of elevated levels of small, dense LDL cholesterol,” Lipids, vol. 41, no. 7, pp. 647–654, Jul. 2006.
[20] J. St-Pierre, I. Lemieux, M. C. Vohl, P. Perron, G. Tremblay, J. P. Despr´es, and D. Gaudet, “Contribution of abdominal obesity and hypertriglyceridemia to impaired fasting glucose and coronary artery disease,” Amer. J. Cardiol., vol. 90, no. 1, pp. 15–18, Jul. 2002.
[21] P. Blackburn, I. Lemieux, N. Alm´eras, J. Bergeron, M. Cˆot´e, A. Tremblay, B. Lamarche, and J. P. Despres, "The hypertriglyceridemic waist phenotype versus the national cholesterol education program-adult treatment panel III and international diabetes federation clinical criteria to identify high-risk men with an altered cardiometabolic risk profile,” Metabolism, vol. 58, no. 8, pp. 1123–1130, Aug. 2009.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY