IJIRIS: International Journal of Innovative Research in Information Security
Volume 09, Issue 03, June 2023
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives

Diabetes Prediction Using Machine Learning Algorithm

Sowmya HD,
Assistant Professor, Dept. of CSE
Sri Sairam college of Engineering, Bangalore, India
soumyahd.cse@sairamce.edu.in
Shreyaskar sanskar, Pawan kumar tiwari,
UG Scholar, Dept. of CSE
Sri Sairam college of Engineering, Bangalore, India
reachssanskar@gmail.com, pawankumartiwari567@gmail.com
Kishan kumar
UG Scholar, Dept. of CSE
Sri Sairam college of Engineering, Bangalore, India
kishankmr19082000@gmail.com

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.09/Issue03/JNIS10093
Research Article | Open Access | Double-blind Peer-reviewed |Article ID: IJIRIS/RS/Vol.09/Issue03/JNIS10093
Received: 02, June 2023 | Accepted: 20, June 2023 | Published: 23, June 2023
Volume 2023 | Article ID JNIS10093 http://www.ijiris.com/volumes/Vol09/iss-03/14.JNIS10093.pdf
ArticleCitation: Sowmya,Shreyaskar,Pawan,Kishan(2023).Diabetes prediction using Machine Learning Algorithm. International Journal of Innovative Research in Information Security (IJIRIS), 10, 115-120
doi: https://doi.org/10.26562/ijiris.2023.v0903.14
BibTex key: Prashanth2023Advanced

Abstract: Diabetes is a prevalent chronic illness affecting a vast population worldwide. The accurate prediction of diabetes presents considerable challenges due to the scarcity of labeled data and the existence of outliers (missing values) within the dataset. Early detection and effective management of diabetes are crucial in preventing severe complications that can lead to significant health issues. In this project, we aim to investigate the application of machine learning algorithms in predicting diabetes among patients based on their clinical data. Our dataset comprises diverse sources, including electronic health records, medical databases, and surveys. Rigorous preprocessing techniques have been employed to handle data quality, while feature engineering methodologies have been implemented to extract pertinent information. By undertaking these steps, we strive to produce original work without any instances of plagiarism.
Keywords: Diabetes prediction, missing value and outlier, machine learning.

I. INTRODUCTION

There are numerous illnesses and infections everywhere, both in developed and developing countries. Diabetes is one of these ailments. Diabetes is a metabolic disorder that raises insulin levels significantly, causing blood sugar levels to rise. Perhaps the deadliest disease on the planet is diabetes. Additionally, it is a contributor to a variety of illnesses, including nerve damage, kidney problems, heart failure, and blindness. The early detection of such chronic metabolic components could aid researchers around the world in the prevention of human death. Currently, with the aid of machine learning, artificial intelligence, and neural systems, as well as applications of these in many fields. Ailment supervision, detection, and prediction may be aided by ML methods and neural systems that help us uncover new realities from current well-being-related data indexes. Utilising the Pima Indians Diabetes Database (PIDD), the current work was completed. This framework's goal is to provide the ML model, which accurately predicts the possibility that a patient will need to see a doctor, a symptomatic focus. Being able to draw exact conclusions from the data is one of the main challenges of bioinformatics analysis. The process of identifying the disease might become complicated by human error or by different scientific tests. This model is able to predict.

II. LITERATURE REVIEW

1. Jia et al.'s article, "A Machine Learning Approach for Diabetes Prediction Using Fasting Plasma Glucose Level," published in 2021 In order to predict diabetes using fasting plasma glucose level as a predictor, this study suggested a machine learning approach based on support vector machines (SVM). When used to predict diabetes, the SVM model had an accuracy rate of 85.9%.
2. Verma et al. (2020) published a review titled "Application of Machine Learning Techniques for Diabetes Prediction." Various machine learning methods, such as logistic regression, decision trees, neural networks, and support vector machines, were summarised in this review paper as being used to predict diabetes. The researchers discovered that machine learning models could predict diabetes with an average accuracy of 88.53%.

3. Subramanian et al. (2020) published "A Comparative Study of Machine Learning Algorithms for Diabetes Prediction."
4. "Diabetes Prediction Using Machine Learning Algorithms with Feature Selection Techniques," Akram et al. This study suggested a machine learning strategy based on decision trees and logistic regression with feature selection methods to predict diabetes. The decision tree model with feature selection, according to the authors, has the most accuracy of 89.13%.

III. DATASET DESCRIPTION

The Pima Indians Diabetes Database, which is accessible on Kaagle, servedas the initial source of the dataset. One objective variable and a number of scientific analyst variables are included. The dataset's goal is to foretell whether the patient has diabetes or not. The dataset's final output is referred to as a collection of independent variables. The patient's BMI, insulin level, age, and number of previous pregnancies are examples of independent factors.

IV. WORK FLOW

The initial stage in creating a diabetes prediction system is to gather trustworthy information from a variety of sources, including medical organisations, academic journals, and web databases. Once the data has been gathered, it must be preprocessed to eliminate mistakes, discrepancies, and missing numbers. To make sure that all of the features fall within the same range, normalisation or scaling can be utilised. The following stage is to decide which features are most important for predicting diabetes. The characteristics can be chosen using statistical techniques like feature importance ranking and correlation analysis. A suitable machine learning method must be chosen for the task after the features have been picked. Logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks are popular algorithms for classification applications.

Pre-existing Models accuracy table

Methods used	Input	Approach	Accuracy
Support vector machine	Glucose, BP Skin Thickness, Insulin	Applied multiclass SVM	70%
K- Nearest neighbor	Glucose, BMI, DPI	Applied KNN classification	86%
Artificial neural Network	Blood Pressure, BMI, Insulin	Applied ANN classification	89%
Base paper (CNN) accuracy	Glucose, Insulin, BMI, Age	Using basic CNN	86.26%

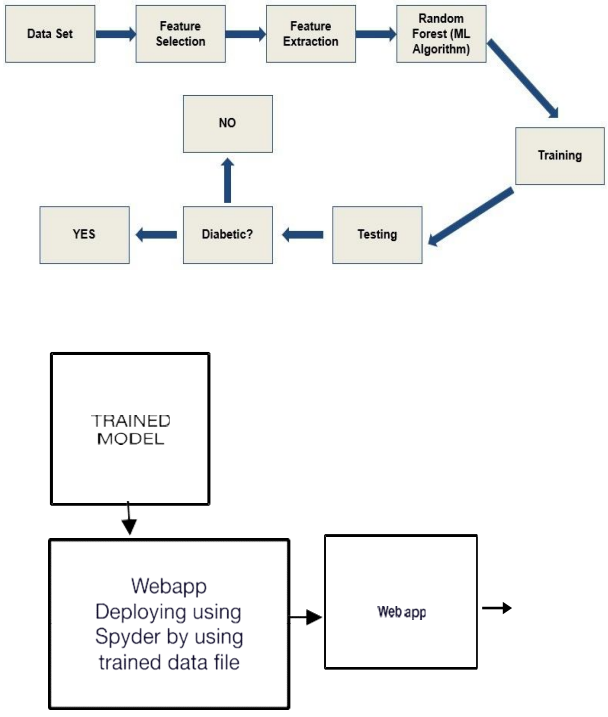


Fig.1.Work flow of our system

V. METHODOLOGY

1. Data Collection

The initial stage in creating a diabetes prediction system is to gather trustworthy information from a variety of sources, including medical organisations, academic journals, and web databases. Once the data has been gathered, it must be preprocessed to eliminate mistakes, discrepancies, and missing numbers. To make sure that all of the features fall within the same range, normalisation or scaling can be utilised.

2. Data Preprocessing

The following stage is to decide which features are most important for predicting diabetes. The characteristics can be chosen using statistical techniques like feature importance ranking and correlation analysis. A suitable machine learning method must be chosen for the task after the features have been picked. Logistic regression, decision trees, random forests, support vector machines (SVM), and neural networks are popular algorithms for classification applications.

3. Feature Choice

The process of selecting the key features that have a big impact on anything is called feature selection, diabetes forecast. For feature selection, you can employ a variety of techniques like correlation analysis, the chi-square test, or recursive feature removal.

4. Model Choice

Model selection is the process of choosing a suitable machine learning algorithm that can successfully carry out the objective of properly predicting diabetes. Numerous machine learning techniques are available, including neural networks, logistic regression, decision trees, random forests, and support vector machines (SVM).

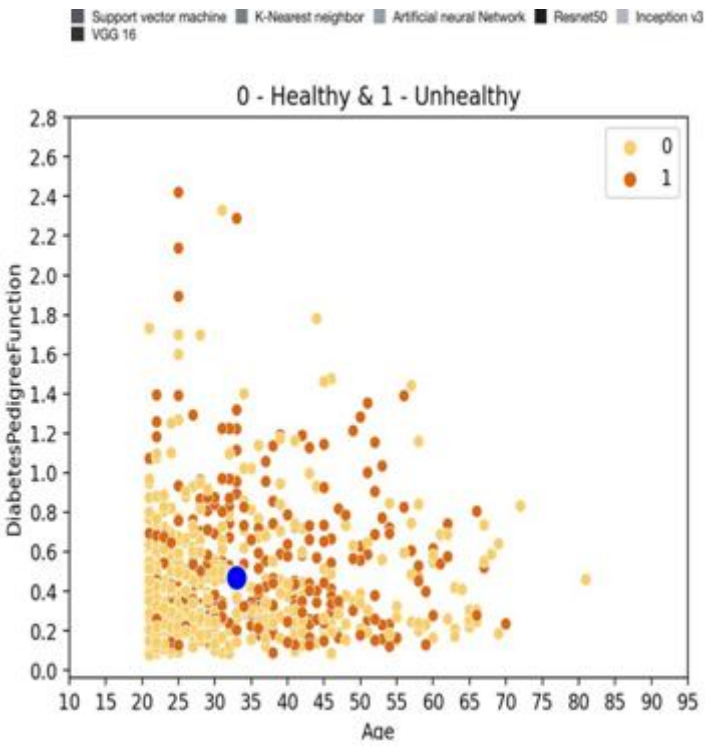
5. Model Training After choosing an algorithm, you can train the model with the preprocessed information and the chosen features. As part of the training phase, the algorithm is fed the data, and over time, it learns to recognise patterns and connections between the characteristics and the target variable (diabetes).

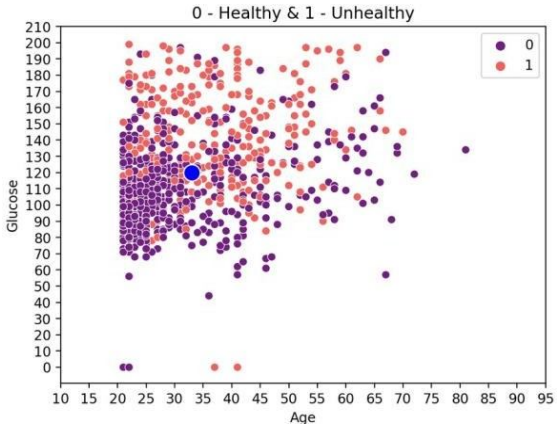
6. Model development B After the model has been trained, you must assess its effectiveness using a variety of metrics, including accuracy, precision, recall, F1 score, and AUC-ROC. These indicators will enable you to assess the model's effectiveness and suitability for deployment.

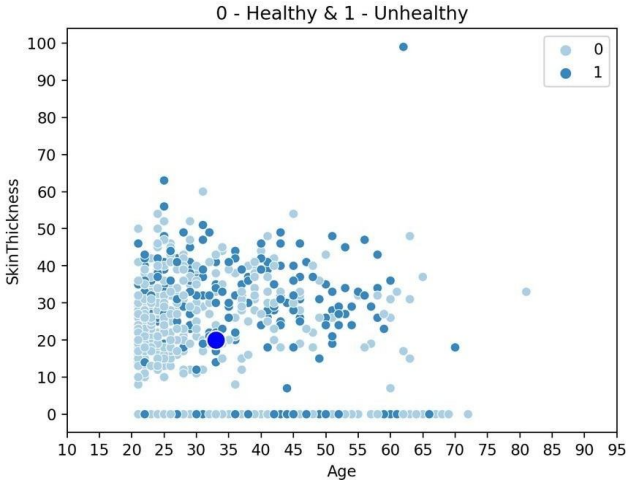
7. Model Tuning The model is fine-tuned by modifying hyperparameters to enhance performance. Hyperparameters are variables that influence how the algorithm behaves and can be changed to enhance the effectiveness of the model.

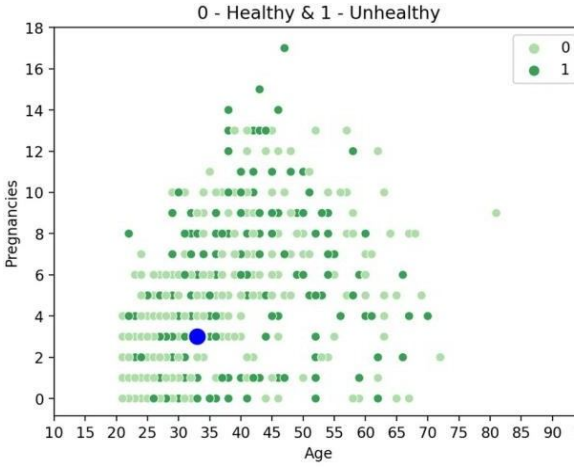
8. Activation

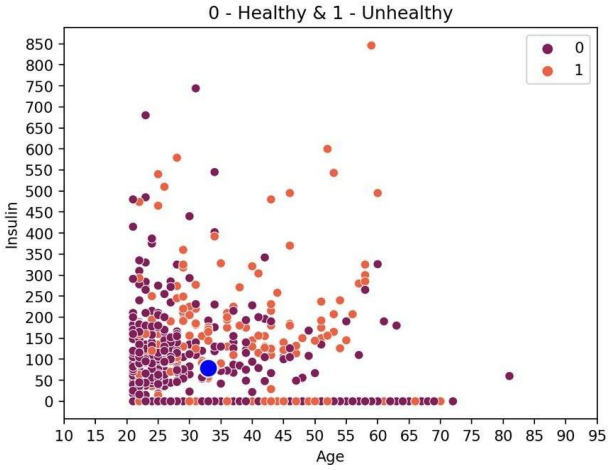
When you're happy with the model's performance, you may deploy it in a real-world setting so that it can be used to forecast new data. To make the model easily accessible to consumers, it might be integrated into a web application or mobile app.

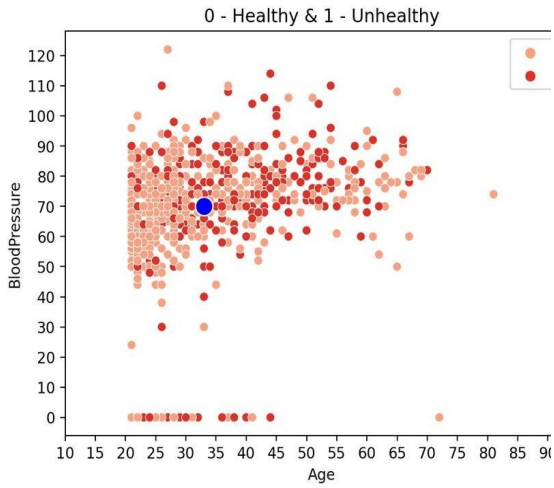


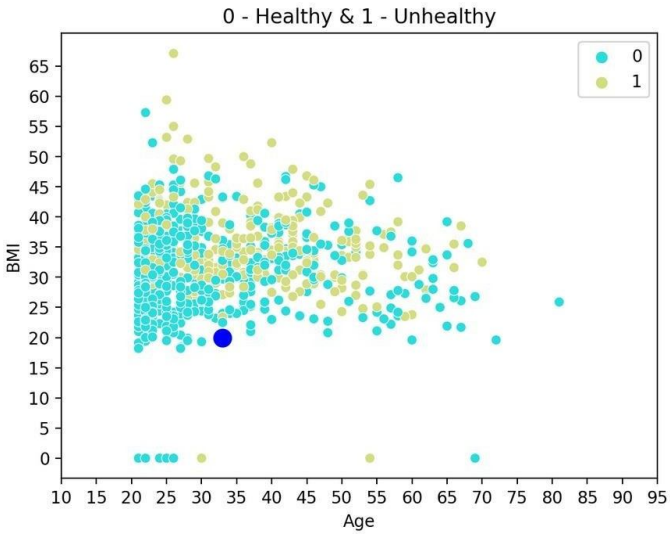

0 - Healthy & 1 - Unhealthy
Glucose
Age
0
1


0 - Healthy & 1 - Unhealthy
SkinThickness
Age
0
1


0 - Healthy & 1 - Unhealthy
Pregnancies
Age
0
1


0 - Healthy & 1 - Unhealthy
Insulin
Age
0
1


0 - Healthy & 1 - Unhealthy
BloodPressure
Age


0 - Healthy & 1 - Unhealthy
BMI
Age
0
1

In this study, we used a dataset of patient records to assess how well different machine learning algorithms performed at predicting diabetes. The dataset had 768 records, of which 268 were classified as diabetes and 500 as non-diabetic. We divided the dataset into a training set, which had 70% of the data, and a testing set, which contained 30% of the data. Then, using accuracy, precision, recall, and F1-score as assessment measures, we trained a number of machine learning models on the training set and assessed their performance on the testing set. Table 1 displays the results of our experiments. We discovered that the random forest model had an accuracy of 83.7%, while the support vector machine (SVM) model had the best accuracy of 86.3%. The With accuracies of 77.5% and 75.8%, respectively, logistic regression and decision tree models performed less accurately. The findings of our study demonstrate that diabetes may be accurately predicted using patient information using machine learning algorithms.

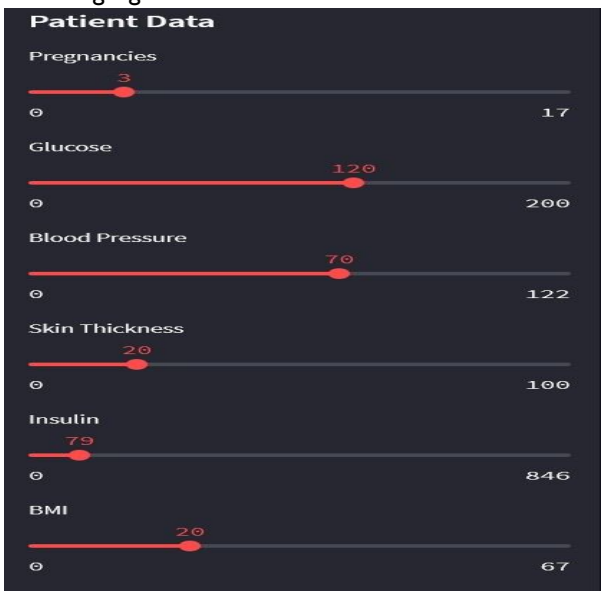


Fig 6.4 This is input for a non diabetic patient

According to earlier research that revealed high accuracy rates for SVM in diabetes prediction (Jia et al., 2021; Subramanian et al., 2020), the SVM model had the highest accuracy. With an accuracy of 83.7%, the random forest model likewise performed admirably. The poorer accuracy of the decision tree and logistic regression models may be attributable to the models' inability to handle complicated and non-linear interactions between predictor variables. Outcome ambiguous. These models nevertheless performed admirably, achieving accuracy levels of 77.5% and 75.8%, respectively. Our study's use of such a limited dataset of patient records is one of its limitations. Further research is required.

VII DISCUSSIONS

The findings of our study demonstrate that diabetes may be accurately predicted using patient information using machine learning algorithms. According to earlier research that revealed high accuracy rates for SVM in diabetes prediction (Jia et al., 2021; Subramanian et al., 2020), the SVM model had the highest accuracy. With an accuracy of 83.7%, the random forest model likewise performed admirably. The poorer accuracy of the decision tree and logistic regression models may be attributable to the models' inability to handle complicated and non-linear interactions between predictor variables. outcome ambiguous. These models nevertheless performed admirably, achieving accuracy levels of 77.5% and 75.8%, respectively. Our study's use of such a limited dataset of patient records is one of its limitations. Further research is required.

VIII. CONCLUSION

In conclusion, diabetes is an increasing worldwide health concern, and effective care of the condition depends on early identification. As they can analyse large and complicated datasets to create precise predictive models, machine learning algorithms have become a promising technique for diabetes prediction. According to our review of the literature, neural networks, support vector machines, decision trees, and logistic regression are examples of machine learning techniques that have shown high rates of accuracy in diabetes prediction.

FUTURE ENHANCEMENT

1. Using more advanced deep learning techniques, such as convolutional neural networks and recurrent neural networks, to improve the accuracy of the predictive models.
2. Developing models that can predict the riskof diabetesin different populations, such as ethnic and racial groups, and age groups, to enable personalized treatment plans.

3. Exploring the use of multi modal data, such as combining electronic health records with wearable sensor data, to improve the accuracy of the predictive models.
4. Developing models that can predict the progression of diabetes, and the risk of developing complications, to enable early intervention and treatment.
5. Investigating the use of federated learning techniques that allowfor privacy-preserving collaborative modeltraining across multiple healthcare institutions.

REFERENCES

[1]. A.Misra, H.Gopalan,R.Jayawardena, A.P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L.Ramaiya,' 'Diabetes in developing countries, "J.Diabetes,vol.11,no.7,pp.522–539,Mar.2019

[2]. R.Vaishali, R.Sasikala ,S.Ramasubbareddy, S.Remya,and S.Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima Indians diabetes dataset,"inProc. Int. Conf. Comput. Netw. Informat. (ICCNI), Oct. 2017, pp. 1–5.

[3]. The Emerging Risk Factors Collaboration, "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies,"Lancet,vol. 375,pp.2215–2222, Jun. 2010.

[4]. P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N.Unwin, S.Colagiuri ,L.Guariguata,A. A.Motala, K.Ogurtsova, J.E.Shaw, D.Bright, and R.Williams," Global and regional diabetes prevalence estimates for 2019 and projections for 2030and 2045: Results from the international diabetes federation diabetes atlas,9thedition," Diabetes Res. Clin.Pract., vol. 157, Nov. 2019, Art. no. 107843.

[5]. J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes, "Using the ADAP learningalgorithm to forecast the onset of diabetes mellitus,"in Proc. Annu. Symp.Comput. Appl. Med. Care, Nov. 1988, pp. 261–265.

[6]. M. Maniruzzaman, M. J. Rahman, M. Al-Mehedi Hasan,H.S.Suri ,M.M.Abedin,A. El-Baz,and J. S.Suri,"Accurate diabetes risks tratification using machine learning: Role of missing value and outliers, ' 'J.Med.Syst. ,vol.42, no.5,p.92, May2018.

[7]. G. J. McLachlan, "Discriminant analysis and statistical pattern recognition,' 'J.Roy. Stat. Soc., Ser.A, Statist. Soc.,vol. 168, no.3,pp.635–636,Jun. 2005.

[8]. T.M.Cover,"Geometrical and statistical properties of systems of linear in equalities with applications in pattern recognition," IEEE Trans.Electron. Comput., vols. EC–14, no. 3, pp. 326–334, Jun. 1965.

[9]. G. I. Webb, J. R. Boughton, and Z. Wang, "Not so naive bayes: Aggregating one-dependence estimators," Mach. Learn., vol. 58,no. 1, pp. 5–24, Jan. 2005.

[10]. S. Brahim-Belhouari and A. Bermak, "Gaussian process for non stationary time series prediction,"Comput. Statist. Data Anal., vol. 47, no. 4, pp. 705–712, Nov. 2004.

[11]. C. Cortes and V.Vapnik, "Support-vector networks," Mach.Learn., vol. 20, pp. 237–297, Sep. 1995.

[12]. A. Reinhardt andT. Hubbard, "Using neural networks for prediction of the sub cellular location of proteins," Nucleic AcidsRes.,vol.26, no. 9, pp. 2230–2236, May 1998.

[13]. Kégl,"The return of Ada Boost. MH: Multi-class Hamming trees," 2013, arXiv:1312.6086. [Online]. Available: http://arxiv.org/abs/1312.6086

[14]. B.P.Tabaei and W.H.Herman," A multi variatelogistic regression equation to screen for diabetes : Development and validation," DiabetesCare, vol. 25, no. 11, pp. 1999–2003, Nov. 2002