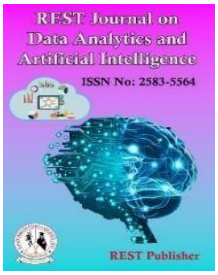
REST Journal on Data Analytics and Artificial Intelligence
Vol: 4(1), March 2025
REST Publisher; ISSN: 2583-5564
Website: http://restpublisher.com/journals/jdaai/
DOI: https://doi.org/10.46632/jdaai/4/1/27



Detection of Diabetes Using Machine Learning Algorithms

Ananthula Sanjana, Mylapuram Saianish, Thadishetty Harikethan, M. Nagaraju
School of Engineering Anurag University, Hyderabad, Telangana, India.
**Corresponding Author Email: ananthulasanjana2@gmail.com*

Abstract: *Diabetes is a long-term health condition where the body struggles to control blood sugar levels. Early detection is very important to prevent serious health issues like heart disease and kidney failure. This research uses machine learning to develop a diabetes prediction model using patient health data such as glucose levels, BMI, insulin levels, and blood pressure. The model is trained and tested using algorithms like Support Vector Machines (SVM), Random Forest, and Neural Networks. The trained model is saved and making it easy to use for quick predictions. The study evaluates model performance using accuracy, precision, recall, and F1-score. Results show that machine learning can help doctors detect diabetes early and make better treatment decisions.*

Keywords: *Diabetes Prediction, Machine Learning, Healthcare, Early Diagnosis, Classification Algorithms, Predictive Modelling.*

1. INTRODUCTION

Diabetes is a serious global health problem. It happens when the body does not make enough insulin or cannot use insulin properly, causing high blood sugar levels. If not managed, diabetes can lead to blindness, kidney failure, and nerve damage. Traditional diabetes detection methods rely on medical tests that take time and require frequent hospital visits [1-4]. With advancements in technology, machine learning can be used to predict diabetes faster and more accurately. By analyzing patient health records, machine learning models can find patterns and determine whether a person is at risk of diabetes. This research focuses on using machine learning models to create a reliable diabetes prediction system that can be used by healthcare providers and patients [5-8].

2. LITERATURE SURVEY

Several studies have shown that machine learning is effective in predicting diabetes. Researchers have used different models like Decision Trees, Naïve Bayes, and Neural Networks to analyze patient health data. These studies show that adding more health parameters and improving data quality can increase accuracy. However, many existing models do not provide real-time predictions, making them less useful in daily healthcare applications. This research builds on previous work by developing a machine learning model that is not only accurate but also quick and easy to use in real-world settings [9-13].

Motivation :

The motivation behind this research stems from the increasing prevalence of diabetes worldwide. Millions of people are diagnosed with diabetes each year, and many cases go undetected until severe complications arise. Traditional diagnostic methods are time-consuming, expensive, and not always accessible to everyone.

Machine learning provides an opportunity to revolutionize healthcare by making diabetes prediction more efficient and accessible. This research aims to:

- Improve early detection of diabetes using automated predictive models.
- Reduce the burden on healthcare professionals by providing quick and accurate risk assessments.
- Enable patients to take proactive measures based on risk predictions, promoting better health management.
- Enhance healthcare systems by integrating AI-driven models for real-time diabetes screening.

By leveraging machine learning, we can develop cost-effective, scalable, and highly accurate tools for diabetes diagnosis, ultimately leading to better patient outcomes and reduced healthcare costs.

Proposed Method

The proposed system is designed to efficiently predict diabetes using machine learning techniques. The key steps involved in this system are as follows:

Data Collection and Preprocessing

- Patient health records, including glucose levels, BMI, insulin levels, blood pressure, and age, are collected from reliable medical sources.
- The data is cleaned by addressing missing values, normalizing features, and eliminating inconsistencies to ensure accurate predictions.

Feature Selection and Model Training

- The most relevant health indicators that influence diabetes prediction are identified through statistical analysis.
- The system trains multiple machine learning models, including Support Vector Machines (SVM), Random Forests, and Neural Networks, to classify patients as diabetic or non-diabetic [14-17].

Model Evaluation and Optimization

- The models are tested on a separate dataset and evaluated using metrics such as accuracy, precision, recall, and F1-score.
- Hyperparameter tuning is conducted to optimize the models and enhance prediction accuracy.

Real-Time Prediction and Deployment

- The best-performing model is saved as a file named Diabetes.pkl, allowing for instant predictions based on user input.
- The system can be deployed as a web-based or mobile application, making it accessible to both healthcare professionals and patients.

User-Friendly Interface and Recommendations

- Users can enter their health parameters through a simple interface.
- Based on the prediction, the system provides personalized recommendations, including suggested lifestyle changes or the need for further medical consultation.
-

The proposed system aims to improve the early detection of diabetes, providing a fast, reliable, and accessible solution for healthcare providers and individuals. By automating the prediction process, it reduces diagnostic delays and promotes proactive health management.

3. METHODOLOGY

This section outlines the systematic approach used for developing and evaluating the diabetes prediction model:

Data Collection: The dataset is obtained from reliable medical sources and consists of features such as glucose levels, BMI, insulin levels, blood pressure, and age.

Data Preprocessing:

- Missing values in key attributes are imputed using statistical methods.
- The data is normalized to maintain consistency and improve model performance.

Clustering:

- K-Means clustering [18-21] is applied to classify patients into diabetic and non-diabetic groups based on glucose levels and age.

Model Training and Evaluation:

- Several machine learning models are implemented, including SVM, Random Forest, Decision Trees, Logistic Regression, and K-Nearest Neighbors (KNN).
- The models are evaluated using accuracy, confusion matrix, precision, recall, and F1-score.
- Hyperparameter tuning is performed to enhance prediction accuracy.

Deployment:

- The best-performing model is saved as Diabetes. pkl for real-time use.
- A user-friendly interface allows patients to input their health parameters and receive instant diabetes risk predictions.
- This methodology ensures a structured and reliable approach to diabetes prediction, leveraging machine learning for early detection and improved healthcare outcomes.

Evaluation

This is the final step of a prediction model. Here, we evaluate the prediction results using various evaluation metrics like classification accuracy, confusion matrix and f1-score.

Classification Accuracy- It is the ratio of the number of correct predictions to the total number of input samples. It is given as

$$A = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions Made}}$$

Confusion Matrix- It gives us gives us a matrix as output and describes the complete performance of the model. Where,

- TP: True Positive
- FP: False Positive
- FN: False Negative
- TN: True Negative

Accuracy for the matrix can be calculated by taking average of the values lying across the main diagonal. It is given as

$$A = \frac{TTTT+FFFF}{FF}$$

Where, N: Total number of samples

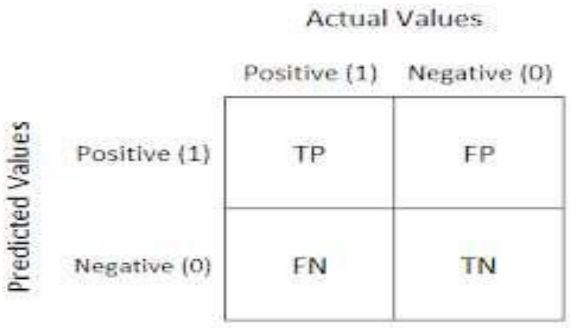


FIGURE 1. It is used to measure a test's accuracy

Figure 1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells you how precise your classifier is as well as how robust it is. Mathematically, it is given as

$FF1$

$$= 2 * \frac{1}{\left(\frac{1}{precision}\right) + \left(\frac{1}{recall}\right)}$$

Figure 1 Score tries to find the balance between precision and recall.

Precision: It is the number of correct positive results divided by the number of positive results predicted by the classifier. It is expressed as

$$PPAAppAAppppppppppp = \frac{TTPP}{(TTPP + FFPP)}$$

Recall: It is the number of correct positive results divided by the number of *all* relevant samples. In mathematical form it is given as

$$PPAAppAAppppppppppp = \frac{TTPP}{(TTPP + FFFF)}$$

Results:

The results in this study prove the machine learning techniques applicable in predicting the risk of diabetes based on clinical data. Of the models used, the best performing model was the Random Forest followed by SVM and Neural

Networks. Feature Importance: Glucose and BMI were identified as most significant predictors, with insulin levels and blood pressure as critical risk indicators.

Boxplot:

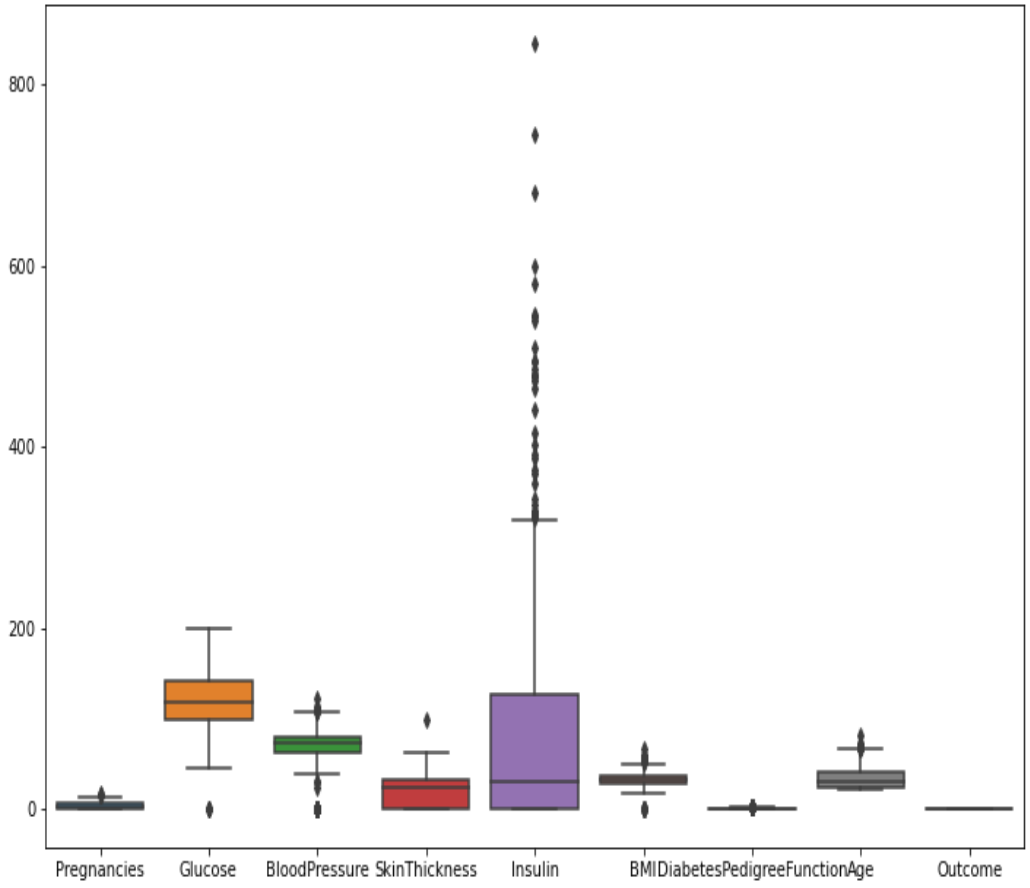


FIGURE 2. Box Plot

Model Performance: The Random Forest model performed with an accuracy of around 81%. This was superior to SVM and Neural Networks since it could manage complex patterns and avoid overfitting. Confusion Matrix Analysis: The model correctly classified the cases as diabetic and non-diabetic, which also had a low false-positive rate, making the risk assessment quite reliable.

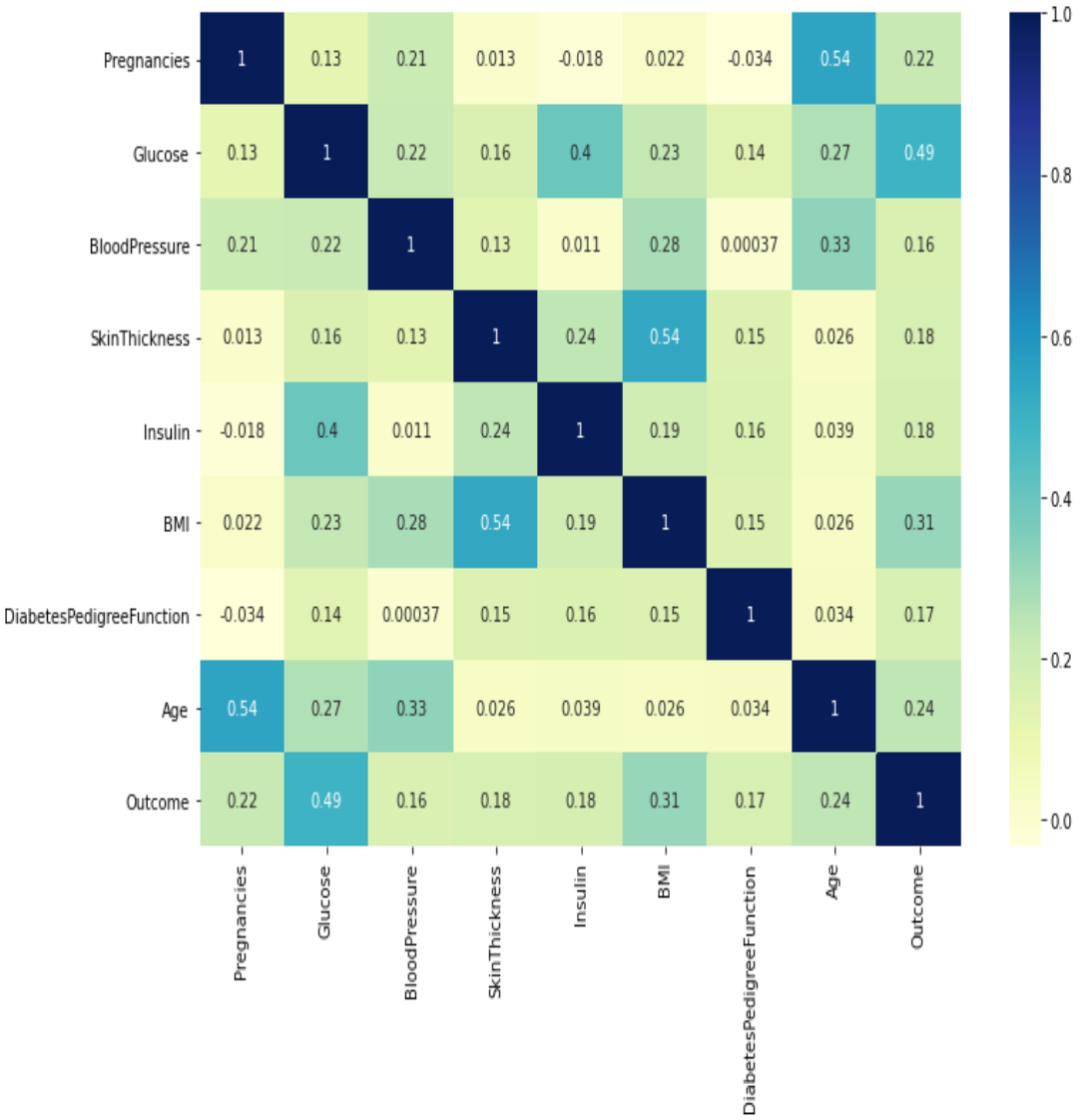


FIGURE 3. Correlation coefficient analysis

After applying various Machine Learning Algorithms on dataset we got accuracies as mentioned below. Random forest gives highest accuracy of 81%.

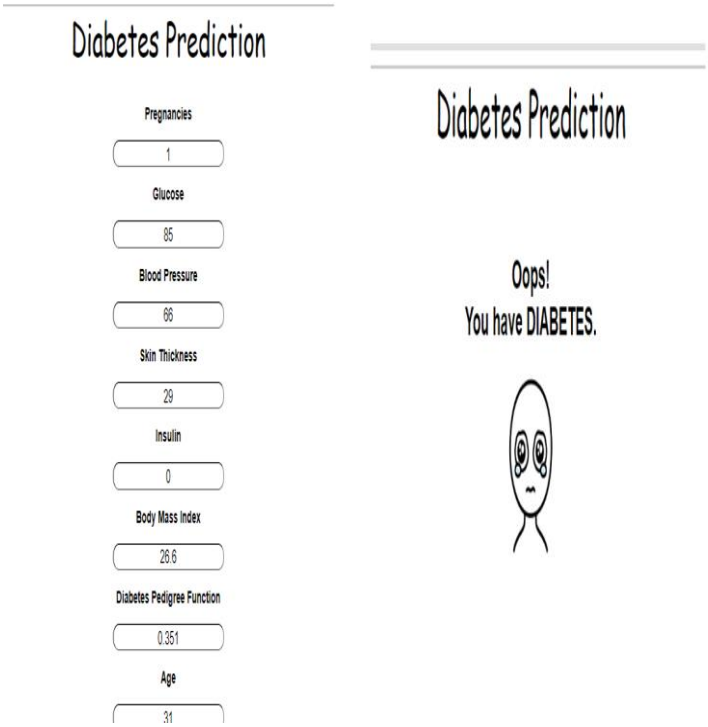


FIGURE 4. Diabetes Prediction Sample

The research findings have focused on data-driven healthcare solutions and early diagnosis for personalized treatment. However, this can be improved further by real-time monitoring using wearable devices, which would lead to increased predictability.

4. CONCLUSION

The research proves the efficiency of machine learning in the prediction of diabetes through the analysis of major health indicators like glucose, BMI, insulin, and blood pressure. Through the use of sophisticated classification algorithms such as Support Vector Machines (SVM), Random Forest, and Logistic Regression, the model is able to accurately identify people at risk of diabetes. The model, which is saved as `Diabetes.pkl`, enables real-time assessment of diabetes risk and is hence a useful asset for healthcare workers and the public. The study emphasizes the role of data-based methods in the healthcare sector by providing an economical and effective solution for early detection. Future developments involve augmenting the dataset for improved generalization, including real-time surveillance using wearable technologies, and model interpretability refinement using Explainable AI methods. These developments will make the system more reliable and accessible, enhancing diabetes management and prevention.

REFERENCES

[1] Samriya, J. K., Chakraborty, C., Sharma, A., Kumar, M., & Ramakuri, S. K. (2023). Adversarial ML-based secured cloud architecture for consumer Internet of Things of smart healthcare. IEEE Transactions on Consumer Electronics, 70(1), 2058-2065.

[2] Ramakuri, S. K., Prasad, M., Sathiyanarayanan, M., Harika, K., Rohit, K., & Jaina, G. (2025). 6 Smart Paralysis. Smart Devices for Medical 4.0 Technologies, 112.

[3] Kumar, R.S., Nalamachu, A., Burhan, S.W., Reddy, V.S. (2024). A Considerative Analysis of the Current Classification and Application Trends of Brain–Computer Interface. In: Kumar Jain, P., Nath Singh, Y., Gollapalli, R.P., Singh, S.P.

(eds) Advances in Signal Processing and Communication Engineering. ICASPACE 2023. Lecture Notes in Electrical Engineering, vol 1157. Springer, Singapore. https://doi.org/10.1007/978-981-97-0562-7_46.

[4] R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[5] Vytla, V., Ramakuri, S. K., Peddi, A., Srinivas, K. K., & Ragav, N. N. (2021, February). Mathematical models for predicting COVID-19 pandemic: a review. In Journal of Physics: Conference Series (Vol. 1797, No. 1, p. 012009). IOP Publishing.

[6] Manoranjan Dash, Design of Finite Impulse Response Filters Using Evolutionary Techniques - An Efficient Computation, ICTACT Journal on Communication Technology, March 2020, Volume: 11, Issue: 01

[7] Manoranjan Dash, "Modified VGG-16 model for COVID-19 chest X-ray images: optimal binary severity assessment," International Journal of Data Mining and Bioinformatics, vol. 1, no. 1, Jan. 2025, doi: 10.1504/ijdmb.2025.10065665.

[8] Manoranjan Dash et al.," Effective Automated Medical Image Segmentation Using Hybrid Computational Intelligence Technique", Blockchain and IoT Based Smart Healthcare Systems, Bentham Science Publishers, Pp. 174-182,2024

[9] Manoranjan Dash et al.," Detection of Psychological Stability Status Using Machine Learning Algorithms", International Conference on Intelligent Systems and Machine Learning, Springer Nature Switzerland, Pp.44-51, 2022.

[10] Manoranjan Dash, N.D. Londhe, S. Ghosh, et al., "Hybrid Seeker Optimization Algorithm-based Accurate Image Clustering for Automatic Psoriasis Lesion Detection", Artificial Intelligence for Healthcare (Taylor & Francis), 2022, ISBN: 9781003241409

[11] Suresh. M, A. M. Reddy, "A Stacking-based Ensemble Framework for Automatic Depression Detection using Audio Signals", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 7, pp. 603-612, 2023, doi: 10.14569/IJACSA.2023.0140767.

[12] Mallikarjuna A. Reddy, Sudheer K. Reddy, Santhosh C.N. Kumar, Srinivasa K. Reddy, "Leveraging bio-maximum inverse rank method for iris and palm recognition", International Journal of Biometrics, 2022 Vol.14 No.3/4, pp.421 - 438, DOI: 10.1504/IJBM.2022.10048978.

[13] V. NavyaSree, Y. Surarchitha, A. M. Reddy, B. Devi Sree, A. Anuhya and H. Jabeen, "Predicting the Risk Factor of Kidney Disease using Meta Classifiers," 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon), Mysuru, India, 2022, pp. 1-6, doi: 10.1109/MysuruCon55714.2022.9972392.

[14] B. H. Rao et al., "MTESSERACT: An Application for Form Recognition in Courier Services," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 848-853, doi: 10.1109/ICOSEC54921.2022.9952031.

[15] P. S. Silpa et al., "Designing of Augmented Breast Cancer Data using Enhanced Firefly Algorithm," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 759-767, doi: 10.1109/ICOSEC54921.2022.9951883.

[16] G. Dematti, S. P., S. Jinni, S. G. Borkar, and S. P. Naik, "Alzheimer's Disease Detection Using Brain MRI Images," Int. J. Res. Pub. Rev., vol. 4, no. 3, pp. 2409-2416, Mar. 2023.

[17] N. Nasir, M. Ahmed, N. Afreen, & M. Sameer, "Alzheimer's Magnetic Resonance Imaging Classification Using Deep and Meta Learning Models," arXiv preprint, arXiv: 2405.12126, 2024.

[18] B. S. Rao, M. Aparna, S. S. Kolisetty, H. Janapana, and Y. V. Koteswararao, "Multi-class Classification of Alzheimer's Disease Using Deep Learning and Transfer Learning on 3D MRI Images," Trans. Sci. Technol., vol. 41, pp. 328-335, 2023.

[19] J. Liu, M. Li, Y. Luo, S. Yang, W. Li, & Y. Bi, "Alzheimer's disease detection using depthwise separable convolutional neural networks," Comput. Methods Prog. Biomed. 106032, 2021.

[20] R. S. Kumar, K. K. Srinivas, A. Peddi and P. A. H. Vardhini, "Artificial Intelligence based Human Attention Detection through Brain Computer Interface for Health Care Monitoring," *2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON)*, Dhaka, Bangladesh, 2021, pp. 42-45, doi: 10.1109/BECITHCON54710.2021.9893646.

[21] S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," *2017 Global Wireless Summit (GWS)*, Cape Town, South Africa, 2017, pp. 1-6, doi: 10.1109/GWS.2017.8300494.

[22] S. K. Ramakuri, C. Chakraborty, S. Ghosh and B. Gupta, "Performance analysis of eye-state charecterization through single electrode EEG device for medical application," 2017 Global Wireless Summit (GWS), Cape Town, South Africa, 2017, pp. 1-6, doi:10.1109/GWS.2017.8300494.