Evaluation of sequential feature selection in improving the K-nearest neighbor classifier for diabetes prediction

Rajkumar Govindarajan[1], Vidhyashree Balaji[1], Jayanthi Arumugam[2], Tsehay Admassu Assegie[3], Radha Mothukuri[4]
[1]Department of Computer Science and Engineering (Data Science),
Madanapalle Institute of Technology and Science, Madanapalle, India
[2]Department of Computer Science and Engineering, Velammal Engineering College (affiliated to Anna University), Chennai, India
[3]Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia
[4]Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

Article Info

Article history:

Received May 23, 2023
Revised Dec 4, 2023
Accepted Dec 14, 2023

Keywords:

Automated diagnosis
Machine learning
Optimized model
Personalized diagnosis
Predictive modeling

ABSTRACT

The K-nearest neighbor (KNN) classifier employs distance metrics to measure the distance between the test instance and the samples used in training. With smaller samples, the KNN classifier achieves higher accuracy with low computational time. However, computing the distance between the test instance and all training samples to determine the class of the test instance requires higher computational time for a high-dimensional dataset. This research employs sequential feature selection (SFS) to select the optimal feature for diabetes prediction while reducing the computational time complexity of the KNN classifier. The KNN classifier showed effectiveness with an accuracy rate of 84.41% with nine features. The performance of the KNN improves by 2.6% when trained on the optimal features selected with the SFS. The result revealed glucose level, blood pressure (BP), skin thickness (ST), diabetes pedigree function (DPF), age, and body mass index (BMI) as the most representative features in diabetes prediction. The KNN classifier gives higher accuracy with these features. However, insulin and the number of times a woman is pregnant do not show a significant effect on the KNN classifier.

Corresponding Author:

Tsehay Admassu Assegie
Department of Computer Science, College of Engineering & Technology, Injibara University
Injibara, Ethiopia
Email: tsehayadmassu2006@gmail.com

1. INTRODUCTION

Machine learning models (MLM) have shown effectiveness in diabetes diagnosis by predicting the pattern within a larger dataset [1]. However, developing an accurate classifier for diabetes prediction remained challenging due to the overlapping features and presence of noise in the dataset. The selection of the MLM model for predicting diabetes requires careful consideration of the model (classifier) behavior.

Diabetes disease has no cure; as such, it should be detected early to reduce the complications caused later. As preventative and treatment measures, MLM has been widely researched and different models have been developed. For instance, Alghamdi [2] has developed an extreme gradient boosting (XGBoost) model for the early prediction of diabetes. The evaluation of the performance of the XGBoost showed that the model achieved an accuracy of 89% for predicting diabetes disease at an early stage. Even though the study has shown that the XGBoost can be effectively used for diabetes prevention and treatment, the developed model has scope for improvement in terms of predictive accuracy.

Similarly, Saxena *et al.* [3] studied the performance of four supervised learning models for diabetes prediction. The study compared the performance of K-nearest neighbor (KNN), random forest (RF), multilayer perceptron (MLP), and decision tree (DT) models for diabetes prediction. The result of the performance evaluation showed that the RF outperforms all three models with an accuracy score of 79.83% while the KNN model scored 78.58 % accuracy for diabetes prediction.

Kakoly *et al.* [4] suggested that the KNN model showed 82.2%, effectiveness for diabetes prediction, the model has the following issues [5]. Firstly, finding the number of nearest neighbors (K value), the distance between the test instances, and all training samples consumes time. Secondly, noisy dataset, with overlapping observations the distance between different training samples and the test instance can be the same and poses difficulty in choosing the majority class. Thirdly, choosing the optimal value of the number of nearest neighbors (K value) is challenging as the KNN model gives different accuracy for different K value.

Machine learning plays an essential part in the healthcare industry by providing ease to healthcare professionals to analyze and diagnose medical data [6]–[10]. Correspondingly, the application of machine learning algorithms to the prediction of diabetes and the reduction of the complexity due to diabetes has become one of the most widely researched areas in recent years. With this regard, Chang *et al.* [11] investigated the performance of different MLM in the treatment and their significance as preventative measures for diabetes. The study assessed the performance of four machine-learning models such as DT, RF, KNN, and logistic regression (LR). The experiment result showed that RF significantly outperformed the others, achieving an accuracy of 82.26%. While the KNN model achieves 80.55% accuracy for diabetes prediction.

Additionally, recent research has shown major development in the application of MLM for predicting diabetes with higher accuracy. One of the widely researched areas in the prediction of diabetes disease with a machine-learning algorithm is feature selection. Feature selection improves the performance of MLM for predicting diabetes [12]. Because with feature selection, noisy and redundant features are removed and the model prediction accuracy improves when the model is trained with relevant features. The experimental result of the study showed that RF score accuracy of 84.1% for diabetes prediction. Even though the study suggests the model performance improvement with feature selection, there is still scope for improving the accuracy of the model for the accurate prediction of diabetes.

Gupta and Goel [13] further enhanced the performance of the KNN model for diabetes prediction by finding the optimal K value for selecting the majority class of the predicted instance. The study showed that the KNN model was 87.01% effective in predicting diabetes. The study also suggested that the performance of the KNN model varies with the number of instances considered for a majority vote by the KNN model. The KNN achieved the highest score with a K value of 4. Correspondingly, Lai *et al.* [14] compared the effectiveness of the DT and KNN model for predicting the presence of diabetes. The comparison between the KNN and DT model shows that KNN has produced better results compared to the DT in terms of predictive accuracy.

Several studies have also applied the KNN model to the prediction of gestational diabetes [15], [16]. The study employed the Public Investment Management Assessment (PIMA) Indian diabetes dataset one of the most commonly employed diabetes datasets for machine learning. The investigation of the performance of LR, MLP, support vector machine, and RF model reveals that the support vector machine outperforms the other models with an accuracy of 87.26%. Additionally, Assegie *et al.* [17] applied shapley additive explanation, a local interpretable model explanation for extracting human understandable insights from XGBoost model prediction. The developed model implemented the XGBoost algorithm for diabetes prediction. Similarly, Haq *et al.* [18] applied ensemble learning methods to diabetes prediction. The researchers conducted an empirical study on the earlier predictability of diabetes with the ensemble learners. The result demonstrates that ensemble learning methods achieve an accuracy of 97% in the early prediction of diabetes. Similarly, Chatterjee *et al.* [19] highlighted the highest performance of the machine-learning model with 79.04%. The study employed the PIMA Indian diabetes dataset. However, the accuracy is different from the previous study [18].

This research aims to develop a KNN model to predict the probability of an individual developing diabetes. The KNN model can be used for early prediction of diabetes, risk assessment, and improving the effectiveness of diabetes prevention and treatment. It can assist healthcare professionals and researchers in predicting diabetes and targeting interventions for diabetes patients. By improving the accuracy of the KNN model for identifying individuals at risk of developing diabetes, the KNN model personalizes diabetes prevention and treatment strategies. Ultimately, reduces the burden on individuals and society as well as lower healthcare costs associated with diabetes treatment. The novelties and contributions of the proposed MLM are: i) evaluation of sequential feature selection (SFS) in reducing the computational time for calculating the distance between the test instance and all training samples, ii) optimizing the dataset by removing the non-informative features in the diabetes dataset, iii) hyperparameter optimization for KNN and demonstration of improvement in accuracy by 2.6%, iv) computation of performance of the KNN model's accuracy and area under the curve of the KNN model for diabetes prediction, and v) benchmarking the proposed MLM with the literature. The rest of the paper is organized as: section 2 contains the method, section 3 describes the result obtained providing the comparison of the KNN classifier performance between the original and feature selected dataset, and section 4 presents the conclusion.

2. METHOD

This study employed the PIMA Indians diabetes dataset, collected from the Kaggle data repository previously studied in [20]–[25] for diabetes prediction. The National Institute of Diabetes, Digestive, and Kidney Diseases provided the source data for this dataset. The dataset's goal is to diagnose whether or not a patient has diabetes based on certain diagnostic metrics provided in the dataset. The dataset includes the following measurements and ranges of clinical test results and physical characteristics. Age (numeric, [21–81 years]), number of times pregnant (continuous 0-17), glucose level (continuous 0-199 mmHg), blood pressure (numerical, [0–199]), skin thickness (numerical, [0–99]), insulin (numerical, [0–846]), body mass index (numerical, [0–67.1]), and diabetes pedigree function (DPF) or the probability of patient getting diabetes due to hereditary factor (numerical, [0.078–2.42]). Table 1 demonstrates the descriptive statistics for the diabetes features employed in developing the KNN model.

Table 1. The descriptive statics of the diabetes features

Variable	Min	Max	STD
Pregnancies	0.00	17.00	3.36
Glucose	0.00	199.0	31.97
Blood pressure	0.00	122.0	19.35
Skin thickness	0.00	99.0	15.95
Insulin	0.00	846	115.24
BMI	0.00	67.10	7.88
DPF	0.078	2.42	0.33
Age	21	81	11.76

3. RESULTS AND DISCUSSION

This section explains the results of the research and provides a comprehensive discussion of the results obtained in the research. Firstly, the most representative features of the diabetes dataset are analyzed with the help of SFS demonstrated in Figure 1. Then the plot shown in Figure 1 visualizes the effect of each feature on the KNN model's (classifier's) diabetes prediction. Based on the features selected, the KNN model is trained on the features that produced better accuracy for the KNN model with 5-fold cross-validation.

Determination of the optimal K value, which provides the highest accuracy score, is important for improving the performance of the KNN model in diabetes prediction. To determine the optimal K value of the proposed KNN model the accuracy score plot over the number of K neighbors is demonstrated in Figure 2. As indicated in Figure 2, the accuracy of the KNN model significantly varies for K values (1-200). Moreover, the variation is different for different K values as shown in the standard deviation plot around each K value. For K values above 3, the variability is very high compared to K values below 3. The model achieves the highest accuracy of 84.41% at K=3.

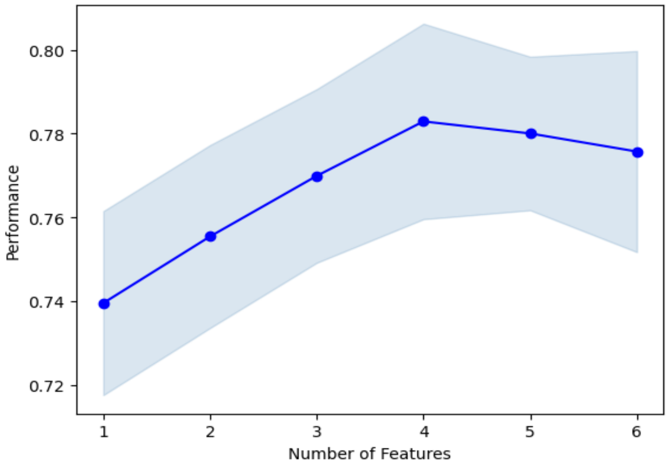


Figure 1. Determination of the optimal number of features with 5-fold cross validation

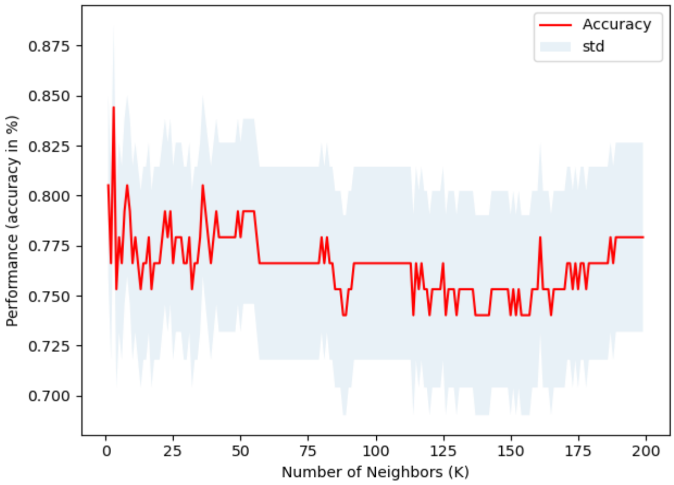


Figure 2. Determination of the optimal value of K

The learning curve, revealed in Figure 3 indicates that as the size of the training set increases, the training score curve and the cross-validation score curve converge. The cross-validation accuracy increases as we add more training data. Thus, adding training data is useful in this case. Since the training score is very accurate, this indicates low bias and high variance. Therefore, this model also begins overfitting the data because the cross-validation score is relatively lower and increases very slowly as the size of the training set increases. The learning curve is a great diagnostic tool to determine bias and variance in a KNN model.

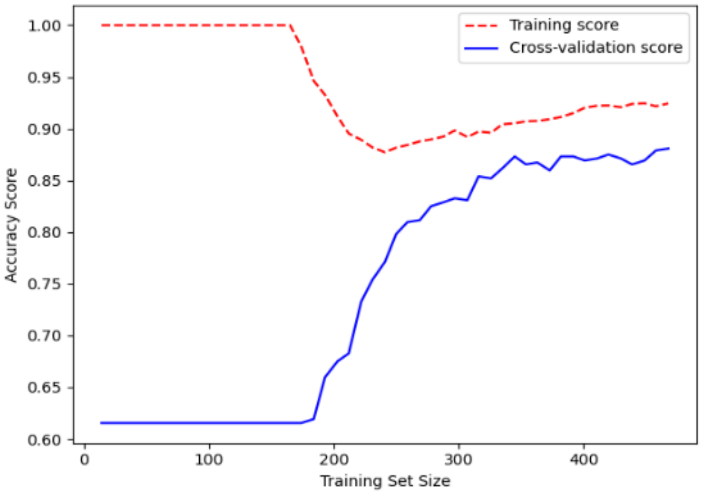


Figure 3. Determination of training samples with 5-fold cross validation

Table 2 validates the KNN model 3-fold cross-validation score on varying numbers of diabetes. As demonstrated in Table 1, the KNN model can achieve across validation accuracy of 68.24% with glucose alone as a predictor feature. Correspondingly, the model’s cross-validation score tends to improve with an increase in body mass index (BMI) as a predictor feature for diabetes. Likewise, the cross-validation accuracy improves with the addition of DPF. However, the cross-validation score remained constant and gradually decreased with the addition of non-significant features namely the insulin, and the number of times a woman is pregnant (pregnancies). After determining the straining sample, size that could give higher possible cross-validation accuracy (692 which is 90% of the 768 diabetes samples) the model is trained on 692 samples. Figure 4 reveals the receiver operating characteristics curve of the KNN model for diabetes prediction. The model scored 0.82 area under the curve for the test set. Table 2 reveals the cross-validation accuracy score and standard error of the KNN model on the original feature set. The researchers employed a pandas data frame to visualize the feature subset the cross-validation score and the standard error. The get_metric_dict method of the sequential feature selector object is used to visualize the selected feature subset. The standard errors of the cross-validation scores, demonstrate the standard error.

Table 2. The performance of KNN on different diabetes features

Variable	Cross validation score	Standard deviation error
Glucose	68.24	0.027
Glucose, BMI	70.95	0.018
Glucose, BMI, DPF	71.41	0.016
Glucose, BMI, DPF, Age	71.49	0.021
Glucose, BMI, DPF, Age, Pregnancies	72.01	0.022
Glucose, BMI, DPF, Age, Pregnancies, ST	72.65	0.018
Glucose, BMI, DPF, Age, Pregnancies, ST, BP, Insulin	72.65	0.015

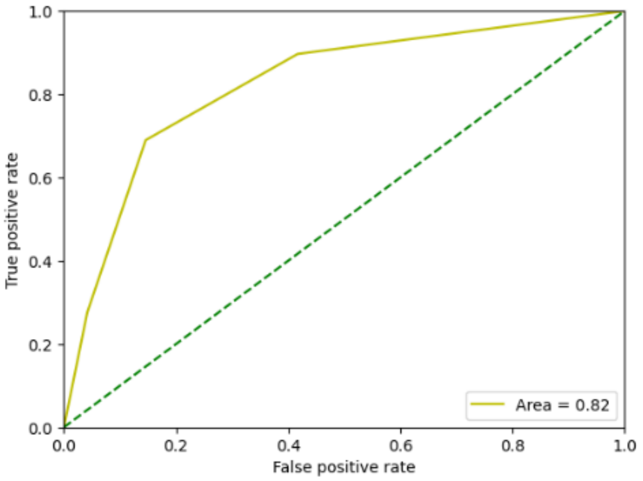


Figure 4. Receiver operating characteristic curve of KNN model

4. CONCLUSION

In conclusion, the SFS technique significantly improved the performance of the KNN classifier for diabetes prediction. The result showed that the use of SFS reduced the number of features required for accurate prediction of diabetes with clinical test results. The study provides additional evidence for the positive association between blood glucose levels and diabetes. The results indicate that polyuria, polydipsia, and sudden weight loss, in particular, are important factors for the prediction of diabetes with KNN models. In contrast with alopecia, the elemental contents in urine and age failed to predict the risk of diabetes due to the low prediction rate when these features are used in the training phase of the KNN. This study highlights the potential of the KNN classifier in early diabetes prediction with clinical test results. Future studies are required to verify the correlation of features in the diabetes dataset to the risk of getting diabetes. It should be pointed out that clinical test result information provided by healthcare experts may also help in revealing the mechanistic relationship between the features of diabetes. Moreover, with pre-processing such as feature selection and hyperparameter optimization through cross-validation, the performance of the KNN model can be improved while reducing the computational time required to calculate the distance between the test instance and all of the training samples. Overall, the result implies that Overall, this SFS can be a useful approach for improving the accuracy and efficiency of MLM in healthcare applications such as diabetes prediction.

REFERENCES

[1] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics*, vol. 2, pp. 1–17, 2022, doi: 10.3389/fbinf.2022.927312.

[2] T. Alghamdi, "Prediction of Diabetes Complications Using Computational Intelligence Techniques," *Applied Sciences*, vol. 13, no. 5, pp. 1–17, 2023, doi: 10.3390/app13053030.

[3] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–11, 2022, doi: 10.1155/2022/3820360.

[4] I. J. Kakoly, M. R. Hoque, and N. Hasan, "Data-Driven Diabetes Risk Factor Prediction Using Machine Learning Algorithms with Feature Selection Technique," *Sustainability*, vol. 15, no. 6, pp. 1–15, 2023, doi: 10.3390/su15064930.

[5] H. Saadatfar, S. Khosravi, J. H. Joloudari, A. Mosavi, and S. Shamshirband, "A new k-nearest neighbors classifier for big data

based on efficient data pruning," *Mathematics*, vol. 8, no. 2, pp. 1–12, 2020, doi: 10.3390/math8020286.

[6] M. Primavera, C. Giannini, and F. Chiarelli, "Prediction and Prevention of Type 1 Diabetes," *Frontiers in Endocrinology*, vol. 11, pp. 1–9, 2020, doi: 10.3389/fendo.2020.00248.

[7] S. S. Bhat, V. Selvam, G. A. Ansari, M. D. Ansari, and M. H. Rahman, "Prevalence and early prediction of diabetes using machine learning in north kashmir: a case study of district bandipora," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, 2022, doi: 10.1155/2022/2789760.

[8] N. Arora, A. Singh, M. Z. N. Al-Dabagh, and S. K. Maitra, "A Novel Architecture for Diabetes Patients' Prediction Using K - Means Clustering and SVM," *Mathematical Problems in Engineering*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/4815521.

[9] N. P. Miriyala, R. L. Kottapalli, G. P. Miriyala, G. Lorenzini, C. Ganteda, and V. A. Bhogapurapu, "Diagnostic Analysis of Diabetes Mellitus Using Machine Learning Approach," *Revue d'Intelligence Artificielle*, vol. 36, no. 3, pp. 347–352, 2022, doi: 10.18280/ria.360301.

[10] S. M. Teki, K. V. Sriharsha, and M. K. V. Nandimandalam, "A diabetic prediction system based on mean shift clustering," *Ingenierie des Systemes d'Information*, vol. 26, no. 2, pp. 231–235, 2021, doi: 10.18280/isi.260210.

[11] V. Chang, M. A. Ganatra, K. Hall, L. Golightly, and Q. A. Xu, "An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators," *Healthcare Analytics*, vol. 2, pp. 1–14, 2022, doi: 10.1016/j.health.2022.100118.

[12] S. Raghavendra and J. Santosh Kumar, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 1, pp. 353–359, 2020, doi: 10.11591/ijece.v10i1.pp353-359.

[13] S. C. Gupta and N. Goel, "Performance enhancement of diabetes prediction by finding optimum K for KNN classifier with feature selection method," in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, pp. 980–986, doi: 10.1109/ICSSIT48917.2020.9214129.

[14] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, no. 1, pp. 1–9, 2019, doi: 10.1186/s12902-019-0436-6.

[15] L. Zhang and M. Liu, "Analysis of Diabetes Disease Risk Prediction and Diabetes Medication Pattern Based on Data Mining," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–9, 2022, doi: 10.1155/2022/2665339.

[16] G. Parthiban and S. K. Srivatsa, "Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients," *International Journal of Applied Information Systems*, vol. 3, no. 7, pp. 25–30, 2012, doi: 10.5120/ijais12-450593.

[17] T. A. Assegie, T. Karpagam, R. Mothukuri, R. L. Tulasi, and M. F. Engidaye, "Extraction of human understandable insight from machine learning model for diabetes prediction," *Bulletin of Electrical Engineering and Informatics*, vol. 11, no. 2, pp. 1126–1133, 2022, doi: 10.11591/eei.v11i2.3391.

[18] A. U. Haq *et al.*, "Intelligent machine learning approach for effective recognition of diabetes in e-healthcare using clinical data," *Sensors*, vol. 20, no. 9, pp. 1–21, 2020, doi: 10.3390/s20092649.

[19] A. Chatterjee, M. W. Gerdes, and S. G. Martinez, "Identification of risk factors associated with obesity and overweight—a machine learning overview," *Sensors*, vol. 20, no. 9, pp. 1–30, 2020, doi: 10.3390/s20092734.

[20] A. D. Jadhav and S. V. Chobe, "Risk Assessment of Cardiovascular Diseases Using kNN and Decision Tree Classifier," *Revue d'Intelligence Artificielle*, vol. 36, no. 1, pp. 155–161, 2022, doi: 10.18280/RIA.360118.

[21] Y. Liu, Z. Yu, and H. Sun, "Prediction Method of Gestational Diabetes Based on Electronic Medical Record Data," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–11, 2021, doi: 10.1155/2021/6672072.

[22] A. Dagliati *et al.*, "Machine Learning Methods to Predict Diabetes Complications," *Journal of Diabetes Science and Technology*, vol. 12, no. 2, pp. 295–302, 2018, doi: 10.1177/1932296817706375.

[23] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," in *Materials Today: Proceedings*, 2021, pp. 1–4, doi: 10.1016/j.matpr.2021.07.196.

[24] T. Mahboob Alam *et al.*, "A model for early prediction of diabetes," *Informatics in Medicine Unlocked*, vol. 16, pp. 1–6, 2019, doi: 10.1016/j.imu.2019.100204.

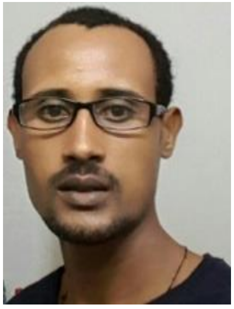
[25] T. Sharma and M. Shah, "A comprehensive review of machine learning techniques on diabetes detection," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, pp. 1–16, 2021, doi: 10.1186/s42492-021-00097-7.

BIOGRAPHIES OF AUTHORS

Rajkumar Govindarajan is currently working as an assistant professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. His research interests include machine learning, data mining, and networking. He can be contacted at email: kumar3544@gmail.com.

Vidhyashree Balaji is currently working as an assistant professor in the Department of Computer Science and Engineering (Data Science) at Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India. Her research interests include machine learning, data mining, and cloud computing. She can be contacted at email: vshree56@gmail.com.

Jayanthi Arumugam is currently working as an assistant professor in the Department of Computer Science and Engineering at Velammal Engineering College, Surapet, Chennai. Her research interests include data mining and machine learning. She can be contacted at email: jayanthiarumugamk@gmail.com.



Tsehay Admassu Assegie has received his M.Sc. in Computer Science from Andhra University, India 2016. He received his B.Sc. in Computer Science from Dilla University, Ethiopia, in 2013. He is currently working as a lecturer in the Department of Computer Science, College of Engineering and Technology, Injibara University, Injibara, Ethiopia. His research includes machine learning, the application of machine learning in healthcare, network security, and software-defined networking. His research has been published in many reputable international journals and international conferences. He is a member of the International Association of Engineers (IAENG). He has reviewed many papers published in different scientific journals. He is an active reviewer of different reputed journals. Recently, Web of Science has verified over 9 peer reviews by him, published in multi-disciplinary digital publishing institute (MDPI) journals. He can be contacted at email: tsehayadmassu2006@gmail.com.

Radha Mothukuri is currently working as an associate professor in the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India. Her research interests include artificial intelligence, machine learning, data mining, and cloud computing. She can be contacted at email: radhahemanth12@gmail.com.