Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 252 (2025) 165–171

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Evolutionary Computing and Mobile Sustainable Networks

Metaheuristic Feature Selection for Diabetes Prediction with P-G-S Approach

Karuppasamy M[a]*, Jansi Rani M[b], Poorani K[c]

[a] *RajaRajeswari College of Engineering, Bengaluru, Karnataka, 560074, India.*
[b] *Mepco Schlenk Engineering College, Sivakasi, Tamilnadu, 626005, India.*
[c] *Kalasalingam Academy of Research and Education, Krishnankoil, Tamilnadu, 626125, India.*
**Corresponding Author*

Abstract

Diabetes mellitus is increasing in large numbers globally. It also leads to various complications ultimately leads to death. Increase in mortality due to diabetic complication is increasing. Early diagnosis of diabetes and its complication leads to decrease in mortality rate. Trending technologies like machine learning and deep learning are much helpful in diagnosing diseases. Advanced prediction models are built to find diseases earlier with efficient algorithms.Metaheuristic approach with particle swarm optimization is utilized. Stacking is one such method which helps in combining various weak algorithms and performs prediction with a meta-classifier. This proposed ParticleSwarm-Gridsearch-Stacking (PGS) approach encompasses the several methods for diabetes prediction and provides the best practices with stacking techniques. The results show that stacking with heterogenous machine learning algorithms with hyperparameter tuning and logistic regression as meta-classifier yields 96.7 % accuracy, 94 % precision and 92.7% recall. This work highlights the importance of feature selection, grid search and stacking which adds overall improvement in machine learning model.

Peer-review under responsibility of the scientific committee of the 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks

Keywords: Diabetes prediction; bagging; boosting; stacking; machine learning algorithms.

1 Introduction

Diabetes is a growing in large numbers due to various reasons. According to IDF report there were 637 million diabetic people at present [1]. Diabetes is caused by high glucose levels in the body. Diabetes is classified into

* Corresponding author. Karuppasamy M
E-mail address: karuppasamy.m1987@gmail.com

1877-0509
Peer-review under responsibility of the scientific committee of the 4th International Conference on Evolutionary Computing and Mobile Sustainable Networks
10.1016/j.procs.2024.12.018

various types depending upon how insulin is utilized by the human body. Type 1 diabetes is called insulin deficient since this case does not produce insulin, so they are called insulin dependent. It is found during childhood and teenages. Type 2 diabetes produce insulin but not absorbed for utilization, hence they are insulin resistant [2]. It is found during adult stages [3]. Gestestional Diabetes is found during pregnancy, in some cases it may disappear after delivery [4]. Causes of diabetes include family history (genetic), physical inactivity, age, BMI and so on.

Diagnosing diabetes at earlier stages reduces the burden of economic cost and prevent from further complications. The hinderence factors that causes complication needs to be identified. The reason for progression of diabetes differs from each. This will increase in the upcoming years [5]. This alarming rate of increase in projection of diabetic numbers highlights the prevalance of diabetic.

Machine learning models are supporting decision making, classification, prognosis, regression for various kinds of medical data. Healthcare data are usually in large numbers which need a automated intervention for result analysis than manual methods. This kind of data also contains noise or missing values. Machine learning performs all kind of preprocessing to classify the data accordingly. It helps to identify the major features contributing to the disease with various feature importnace methods. Machine learning works by preprocessing, dividing the dataset, thereby train and test with various models.

2 Related Work

Le TM et al [6] work used grey wolf optimization and adaptive particle swarm optimization for feature selection and MLP for classification which achieved 97% accuracy. Li X et al [7] combined harmony search, genetic algorithm, particle swarm with kNN classifier and produced 91.6% accuracy.Talari P et al [8] in their work proposed combined feature selection with sequential minimal optimization (SMO) and synthetic minority optimization (SMOTE). This combined approach with bagging decision tree is used for diabetic prediction. Abdollahi J et al [9, 10] in their work used particle swarm optimization for feature selection, while random forest, decision tree and naïve bayes achieved high accuracy with minimum error rate.

Dogru A et al [11] proposed super learner ensemble method with four base learners and SVM as meta learner. Chi-sq for optimal feature selection and hyperparameter tuning with grid search was performed. Tasin I et al [12] used mutual information for feature selection, SMOTE and ADASYN for class imbalance handling and XGBoost classifier and achieved accuracy of 81 %. They validated their performance through LIME and SHAP framework.

Sneha N and Gangil T [13] in their work utilized optimal features for classification, where random forest and decision tree achieved highest specificity while naïve bayes produced high accuracy of 82.30 %. Ali MS et al [14] work used random forest with best parameters for prediction.

Kalagotla et al [15] proposed three methodologies with correlation for feature selection, adaboost on selected features and stacking with MLP, SVM and LR. Khilwani VO et al [16] used SVM, ANN, LR, DT, RF and NB with LR as meta classifier and achieved accuracy of 82.6%. Singh N and Singh P [17] developed NSGA II stacking model with optimization algorithm for base learner and kNN as meta classifier and produced 83.6 % accuracy, 96.1% sensitivity and specificity of 79.9%. Hasan MK et al [18] proposed heterogenous model for classification which works better than homogenous models and achieved accuracy of 93.1%. Tan Y et al [19, 20] proposed genetic algorithm based decision tree, SVM, LR, NB and CNN before classification.

Zhang Z et al [21] in their work used harmony search and optimization techniques for classification and achieved accuracy of 93.04%. Mustofa F et al [22, 23] work compared random forest with two imbalanced dataset and used k fold validation. Rahim MA et al [24] used stacking techniques of various classifiers with logistic regression as meta learner and scored 94% accuracy. Daza A et al [25] work used oversampling and random forest achieved 91.5 % accuracy, while 97% ROC is achieved with stacking.

3 Proposed Work

The proposed work focuses on exploring various ensemble techniques used for diabetes prediction.
The proposed work is classified as

- Data Preprocessing
- Feature Selection
- Machine learning
- Performance Metrics

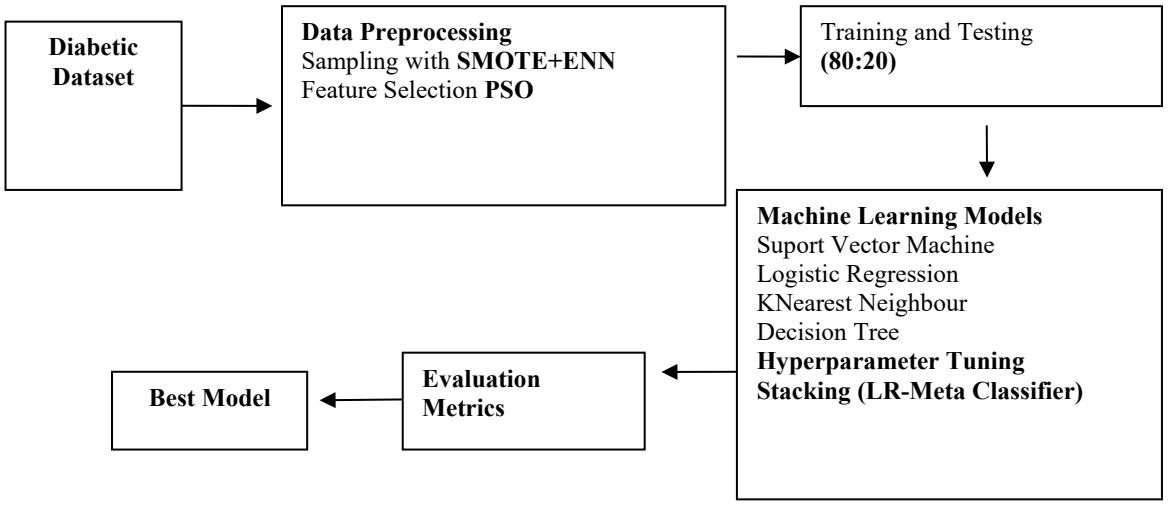


Fig 1: Proposed PGS Work flow

Fig 1 represents the workflow designed for the prediction model. It involves data preprocessing, feature selection and machine learning models.

3.1 Data Preprocessing

Data preprocessing is usually done to improve the performance of any model given to the dataset. The PIMA Indian diabetic dataset is taken for the proposed work. It consists of 768 data with 9 columns. Feature selection involves selecting the major features contributing to the outcome. Table I describe the dataset used for diabetes classification. Normalization with Z-score is performed for diabetic dataset inorder to normalize the all the features to perform further classification.

Table 1: Dataset Description

S.No	Attributes
1	Pregnancy
2	Glucose
3	Diastolic Blood pressure(BP)
4	Skin Thickness
5	Serum Insulin
6	BMI
7	Diabetes pedigree Function
8	Age
9	Outcome

Table 1 represents the dataset taken for the proposed work. This also depicts various dependant and independent variable used. Outcome determines whether the person is diabetic or non-diabetic.

3.2 Feature Selection

Particle Swarm Optimization is a computational method which optimizes a problem iteratively and improve the

candidate solution. Population of candidate solution solves the issue. Each particle's movement is influenced by its local best positions, also move towards best known positions updated by other particles.

Consider the parameters
Initialization;
while reached maximum iteration
Evaluation of objective function;
Constraint balancing with penalty;
Considering each particle with its velocity;
Update particle (best);
Position updation
End while
Result Displayed.

This is the algorithm used to find the best positions using particle swarm optimization technique.

3.3 Machine Learning Models

3.3.1 Support Vector Machine (SVM)

SVM, the most common technique in classification splits the classes with the help of hyperplane. The hyperplane should not exceed the number of input features. Largest margin between decides the best hyperplane for classification.

$$y=wX+b \tag{1}$$

3.3.2 Logistic Regression (LR)

LR, an extended form linear regression solves both classification and regression problems. Probability is used to find the classification outcome. Sigmoid function and its threshold value is used for classification. The positive class and negative class depend on whether the threshold value is greater than or less than 0.5.

$$h\theta(X) = \frac{1}{1+e^{-(\beta 0+\beta 1X)}} \tag{2}$$

Sigmoid function equation is as follows:

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

3.3.3 K Nearest Neighbor (KNN)

kNN is a classification algorithm which utilizes distance metrics to find the nearest neighbour for classification. k indicates the neighbour support for voting. This is also known as lazy learner. Euclidean distance metric is used to solve the classification.

Euclidean distance equation is as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)2} \tag{4}$$

3.3.4 Decision Tree (DT)

Decision Tree uses decision from various trees for classification. DT employs gini index and gini ratio for data splitting. Pruning is used to avoid overfitting.

Equation for decision split is as follows

$$Gini(split) = \sum_{i=1}^{n} \text{pi} * (1 - \text{pi}) \quad (5)$$

4 Hyperparameter Tuning (Grid Search)

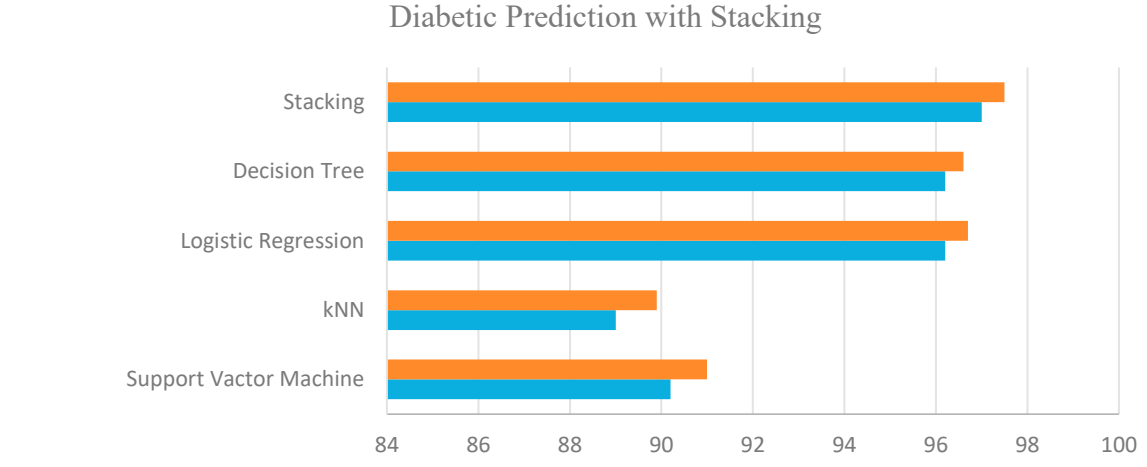
This is done to improve the model performance by choosing the optimal hyperparameters for classification. Gridsearch is one of the hyperparameter tuning methods used for tuning. This involves the list of hyperparameters and performance metric, by working with all possible outcomes. It is effective, but computationallly intensive.

5 Stacking

Stacking allows a training algorithm to ensemble several learning algorithms. It utilizes various trained base learners to predict the outcome. The base learners are stacked to the meta learner for prediction. In this work SVM, LR, kNN and DT are the base learners with logistic regression as meta learner. The prediction with stacking provides much efficient results when compared with traditional methods.

6 Results and Discussion

The proposed work shows the performance analysis of various machine learning algorithms.



	Support Vactor Machine	kNN	Logistic Regression	Decision Tree	Stacking
Accuracy after hyperparameter	91	89.9	96.7	96.6	97.5
Accuracy	90.2	89	96.2	96.2	97

Fig 2: Performance Metrics for Diabetes Prediction

Fig 2 represents results of various machine learning models. It shows that stacking provides better results when comparing single model. This leads to various ensembling techniques for disease prediction. The proposed results prove that feature selection, hyperparameter tuning also contribute to stacking model. This work performs PSO, Gridsearch and stacking to improve the classification performance.

7 Conclusions & Future Work

This proposed work concludes that diabetes prediction model is much improved with adapting particle swarm optimization as a feature selection method and grid search for hyperparameter tuning. The machine learning model is more improved when implementing stacking, eventhough complexity and runtime of the model may slighty be higher than without stacking. Future work aims to explore the complexity in stacking and ways for time reduction.

References

[1] International Diabetes Federation (IDF).
[2] World Health Organization.
[3] Redondo, M. J., Hagopian, W. A., Oram, R., Steck, A. K., Vehik, K., Weedon, M., ... & Dabelea, D. (2020). "The clinical consequences of heterogeneity within and between different diabetes types" *Diabetologia*, *63*, 2040-2048.
[4] Alwash, S. M., McIntyre, H. D., & Mamun, A. (2021). "The association of general obesity, central obesity and visceral body fat with the risk of gestational diabetes mellitus: Evidence from a systematic review and meta-analysis" *Obesity Research & Clinical Practice*, *15*(5), 425-430.
[5]Gregory, G. A., Robinson, T. I., Linklater, S. E., Wang, F., Colagiuri, S., de Beaufort, C., ... & Ogle, G. D. (2022). "Global incidence, prevalence, and mortality of type 1 diabetes in 2021 with projection to 2040: a modelling study" *The lancet Diabetes & endocrinology*, *10*(10), 741-760.
[6] Le, T. M., Vo, T. M., Pham, T. N., & Dao, S. V. T. (2020). "A novel wrapper–based feature selection for early diabetes prediction enhanced with a metaheuristic" *IEEE access*, *9*, 7869-7884.
[7] Li, X., Zhang, J., & Safara, F. (2023). "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm" *Neural processing letters*, *55*(1), 153-169.
[8] Talari, P., N, B., Kaur, G., Alshahrani, H., Al Reshan, M. S., Sulaiman, A., & Shaikh, A. (2024). "Hybrid feature selection and classification technique for early prediction and severity of diabetes type 2" *Plos one*, *19*(1), e0292100.
[9] Poorani, K., & Karuppasamy, M. (2023, April). "Comparative Analysis of Chronic Kidney Disease Prediction Using Supervised Machine Learning Techniques" In *International Conference on Information and Communication Technology for Intelligent Systems* (pp. 87-95). Singapore: Springer Nature Singapore.
[10] Abdollahi, J., & Aref, S. (2024). "Early Prediction of Diabetes Using Feature Selection and Machine Learning Algorithms" *SN Computer Science*, *5*(2), 217.
[11] Doğru, A., Buyrukoğlu, S., & Arı, M. (2023). "A hybrid super ensemble learning model for the early-stage prediction of diabetes risk" *Medical & Biological Engineering & Computing*, *61*(3), 785-797.
[12] Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). "Diabetes prediction using machine learning and explainable AI techniques" *Healthcare Technology Letters*, *10*(1-2), 1-10.
[13] Sneha, N., & Gangil, T. (2019). "Analysis of diabetes mellitus for early prediction using optimal features selection" *Journal of Big data*, *6*(1), 1-19.
[14] Ali, M. S., Islam, M. K., Das, A. A., Duranta, D. U. S., Haque, M. F., & Rahman, M. H. (2023). "A novel approach for best parameters selection and feature engineering to analyze and detect diabetes: Machine learning insights" *BioMed Research International*, *2023*(1), 8583210.
[15] Kalagotla, S. K., Gangashetty, S. V., & Giridhar, K. (2021) "A novel stacking technique for prediction of diabetes" *Computers in Biology and Medicine*, *135*, 104554.
[16] Khilwani, V. O., Gondaliya, V., Patel, S., Hemnani, J., Gandhi, B., & Bharti, S. K. (2021). "Diabetes prediction, using stacking classifier" In *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)* (pp. 1-6). IEEE.
[17] Singh, N., & Singh, P. (2020). "A stacked generalization approach for diagnosis and prediction of type 2 diabetes mellitus" In *Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018* (pp. 559-570). Springer Singapore.
[18] Hasan, M. K., Saeed, R. A., Alsuhibany, S. A., & Abdel-Khalek, S. (2022) "An empirical model to predict the diabetic positive using stacked ensemble approach" *Frontiers in Public Health*, *9*, 792124.
[19] Poorani, K., & Karuppasamy, M. (2022, December). "Analysis of Underlying and Forecasting Factors of Type 1 Diabetes and Prediction of Diabetes Using Machine Learning" In *International Conference on Information and Management Engineering* (pp. 93-100). Singapore: Springer Nature Singapore.
[20] Tan, Y., Chen, H., Zhang, J., Tang, R., & Liu, P. (2022) "Early risk prediction of diabetes based on GA-stacking" *Applied Sciences*, *12*(2), 632.
[21] Zhang, Z., Lu, Y., Ye, M., Huang, W., Jin, L., Zhang, G., ... & Zhu, W. (2024) "A novel evolutionary ensemble prediction model using harmony search and stacking for diabetes diagnosis" *Journal of King Saud University-Computer and Information Sciences*, *36*(1), 101873.

[22] Poorani, K., & Karuppasamy, M. (2023, August) "Information System for Neuropathy Prediction Ensembling Ranking and Ordered Clustering for Diabetic Healthcare Monitoring". In *International Conference on ICT for Sustainable Development* (pp. 483-491). Singapore: Springer Nature Singapore.
[23] Mustofa, F., Safriandono, A. N., Muslikh, A. R., & Setiadi, D. R. I. M. (2023). "Dataset and feature analysis for Diabetes Mellitus classification using random forest" *Journal of Computing Theories and Applications (JCTA)*, *1*(1), 41-49.
[24] Rahim, M. A., Hossain, M. A., Hossain, M. N., Shin, J., & Yun, K. S. (2023). "Stacked ensemble-based type-2 diabetes prediction using machine learning techniques" *Annals of Emerging Technologies in Computing (AETiC)*, *7*(1), 30-39.
[25] Daza, A., Sánchez, C. F. P., Apaza-Perez, G., Pinto, J., & Ramos, K. Z. (2024). "Stacking ensemble approach to diagnosing the disease of diabetes" *Informatics in Medicine Unlocked*, *44*, 101427.
[26] Jansi Rani, M., & Devaraj, D. (2019) "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification" *Journal of medical systems*, *43*(8), 235.