ISSN: 2633-4828 Vol. 4 No.2, September, 2022

International Journal of Applied Engineering & Technology

# CLASSIFICATION OF TYPE 2 DIABETES USING ENSEMBLE LEARNING METHODS

**Chandrashekar C M[1] and Dr. Anurag Shrivastava[2]**

[1]Research Scholar and [2]Supervisor, Department of Computer Science and Engineering, Monad University, Hapur (U.P.), 245304, India

[1]chandrucse014@gmail.com

## ABSTRACT

*Type 2 Diabetes (T2D) is a chronic metabolic disorder that affects millions of people worldwide, making early diagnosis crucial for effective treatment and management. Machine learning (ML) techniques have been increasingly utilized to enhance the classification accuracy of T2D by identifying complex patterns in patient data. This paper introduces a novel approach that leverages ensemble learning methods to improve the classification of T2D. By combining multiple classifiers and focusing on optimal feature selection, the proposed method achieves superior performance compared to traditional classification techniques. The effectiveness of the model was validated using multiple real-world healthcare datasets, demonstrating significant improvements in predictive accuracy.*

*Keywords: Type 2 Diabetes, Ensemble Learning, Classification, Machine Learning, Predictive Modelling, Healthcare Analytics*

## 1. INTRODUCTION

The rapid increase in Type 2 Diabetes (T2D) prevalence has placed a significant burden on global healthcare systems. T2D is characterized by insulin resistance and chronic hyperglycemia, leading to severe complications such as cardiovascular diseases, neuropathy, and kidney failure if not managed properly. Early diagnosis and accurate classification of T2D are essential to preventing these complications and improving patient outcomes. However, traditional diagnostic methods, which often rely on linear models and expert knowledge, may not be sufficient to capture the complex, non-linear relationships inherent in patient data. Machine learning (ML) offers a powerful alternative, enabling the analysis of large datasets to identify subtle patterns that may be indicative of T2D. Among the various ML techniques, ensemble learning has gained prominence due to its ability to combine the strengths of multiple models, thereby improving predictive accuracy. This paper explores the application of ensemble learning methods to the classification of T2D, with a particular focus on feature selection and model integration to enhance classification performance. Ensemble learning techniques, such as bagging, boosting, and stacking, have been extensively applied across different domains, including healthcare, to improve classification accuracy. In the context of T2D, Patel et al. (2021) highlighted that ensemble methods consistently outperform individual classifiers by aggregating the predictions of multiple models, thereby reducing the risk of overfitting and increasing model robustness. Feature selection is another critical area in healthcare data analysis. High-dimensional datasets, typical in healthcare, often contain redundant or irrelevant features that can negatively impact model performance. Techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) have been utilized to reduce the dimensionality of datasets, improving both model interpretability and performance. Kaur and Kumari (2018) emphasized the importance of feature selection in enhancing the accuracy of ML models for T2D classification, noting that selecting the most relevant features is crucial for effective model training.

## 2. LITERATURE REVIEW

The classification of Type 2 Diabetes (T2D) has increasingly become a focal point in medical research, with machine learning (ML) methods being utilized to enhance diagnostic accuracy. Ensemble learning, in particular, has shown promise in improving the reliability and performance of models used for disease classification. Ensemble methods combine multiple classifiers to form a robust predictive model, thereby mitigating the weaknesses of individual classifiers. This approach has been extensively explored in the context of healthcare data, where the complexity and high dimensionality of datasets often pose significant challenges to traditional ML models.

### Feature Selection Techniques in Healthcare

One of the critical challenges in healthcare data analysis is feature selection. The presence of noisy, irrelevant, or redundant features can degrade the performance of classification models. Various studies have focused on developing feature selection methods tailored for healthcare datasets. Kaur and Kumari (2018) emphasized the importance of predictive modeling in diabetes using machine learning, noting that appropriate feature selection is crucial for improving model accuracy. Their work underscored that features in healthcare datasets are often correlated, and selecting the most relevant ones can significantly reduce model complexity and enhance interpretability.

### Ensemble Learning Methods

The use of ensemble learning in healthcare has been highlighted by several researchers. For instance, Özçift (2011) demonstrated that random forests, an ensemble learning method, could be effectively used to diagnose cardiac arrhythmias by combining multiple decision trees. The study found that random forests performed better than individual classifiers, particularly in dealing with complex, non-linear relationships in the data. Similarly, Patel et al. (2021) conducted a comprehensive review of hybrid machine learning approaches for predicting T2D, finding that ensemble methods consistently outperformed single models in terms of accuracy and robustness.

### Dimensionality Reduction in Healthcare Data

Dimensionality reduction techniques, such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), have also been widely used to address the challenges posed by high-dimensional healthcare data. Nguyen and Holmes (2019) provided insights into the application of dimensionality reduction techniques in healthcare, advocating for their use to simplify models and reduce computational costs. They argued that reducing the dimensionality of data helps in improving the performance of classifiers by focusing on the most informative features.

### Advancements in Diabetes Classification

The classification of T2D using machine learning has seen significant advancements, with ensemble methods being at the forefront of these developments. Saeed et al. (2021) explored various ensemble learning techniques for predicting T2D and found that these methods could effectively handle the complex interactions between different risk factors. Their study demonstrated that ensemble methods, particularly those incorporating feature selection techniques, could achieve higher accuracy rates compared to traditional classification methods.

### Challenges and Future Directions

Despite the successes of ensemble learning methods, several challenges remain. The selection of the most appropriate classifiers and the integration of feature selection within the ensemble framework are areas that require further exploration. Chou et al. (2004) pointed out that while ensemble learning can improve predictive performance, the process of selecting and tuning individual models for inclusion in the ensemble is non-trivial. Additionally, the computational complexity of ensemble methods, especially when applied to large healthcare datasets, remains a significant hurdle. The literature also suggests a growing interest in the application of deep learning techniques in conjunction with ensemble learning for T2D classification. Liu et al. (2015) highlighted the potential of deep learning in early disease diagnosis, suggesting that future research could explore the integration of deep learning models within an ensemble framework to enhance diagnostic accuracy further.

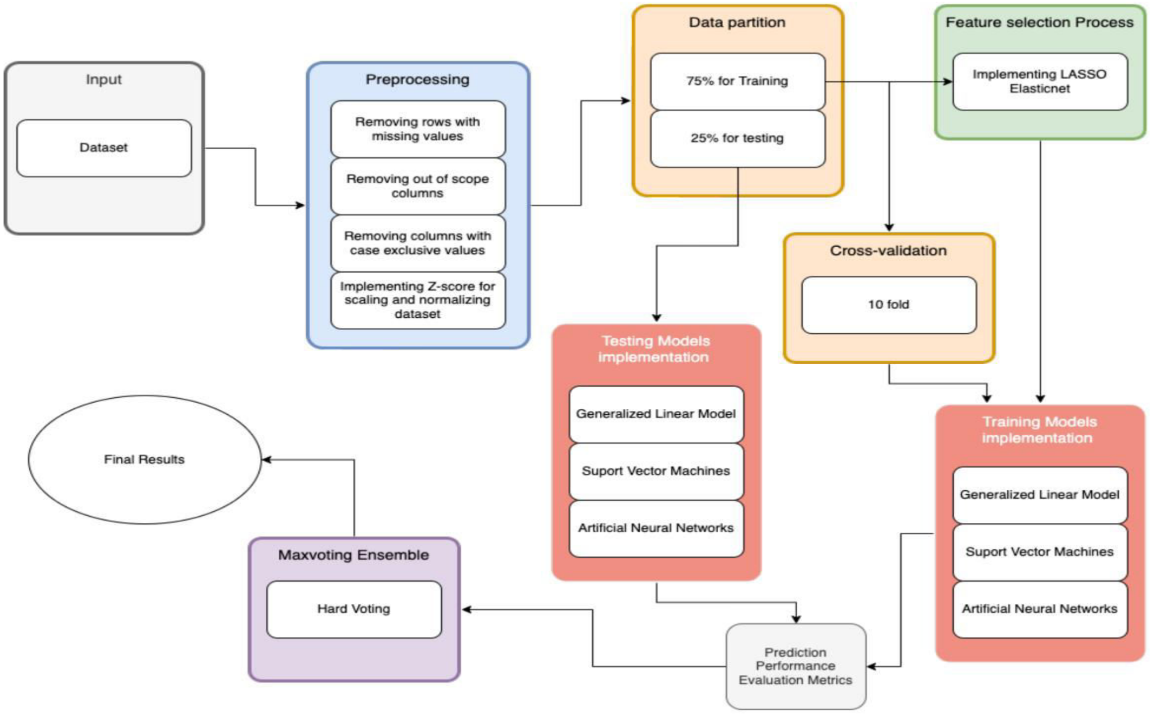

Input
Dataset
Preprocessing
Removing rows with missing values
Removing out of scope columns
Removing columns with case exclusive values
Implementing Z-score for scaling and normalizing dataset
Data partition
75% for Training
25% for testing
Feature selection Process
Implementing LASSO Elasticnet
Cross-validation
10 fold
Testing Models implementation
Generalized Linear Model
Suport Vector Machines
Artificial Neural Networks
Training Models implementation
Generalized Linear Model
Suport Vector Machines
Artificial Neural Networks
Final Results
Maxvoting Ensemble
Hard Voting
Prediction Performance Evaluation Metrics

**Fig.1:** Flowchart of the proposed methodology.

## 3. METHODOLOGY

This study proposes a two-phase approach to T2D classification using ensemble learning methods. The first phase involves the Competitive Ensemble Feature Selection Model (CEFSM), which is designed to identify the most relevant features for classification. The second phase involves the Competitive Ensemble Classification Model for Structured Data using Machine Learning Techniques (CECMSDML), which integrates multiple classifiers to create a robust ensemble model.

### 3.1 Feature Selection

Feature selection is a vital preprocessing step that aims to reduce the dimensionality of the dataset by eliminating irrelevant or redundant features. The CEFSM employs a combination of feature selection techniques, including:

- **Univariate Feature Selection:** This method evaluates each feature individually, selecting those with the highest correlation to the target variable.
- **Correlation-Based Feature Selection:** This method retains features that show a strong correlation with the target variable while discarding those with low correlation.
- **Principal Component Analysis (PCA):** PCA reduces dimensionality by transforming features into a set of linearly uncorrelated components, focusing on those with the highest variance.
- **Recursive Feature Elimination (RFE):** RFE iteratively removes the least important features, focusing on those that contribute most to model accuracy.

### 3.2 Ensemble Classification

The CECMSDML is designed to enhance classification accuracy by integrating multiple classifiers, each contributing to the final prediction based on its performance. The model includes the following classifiers:

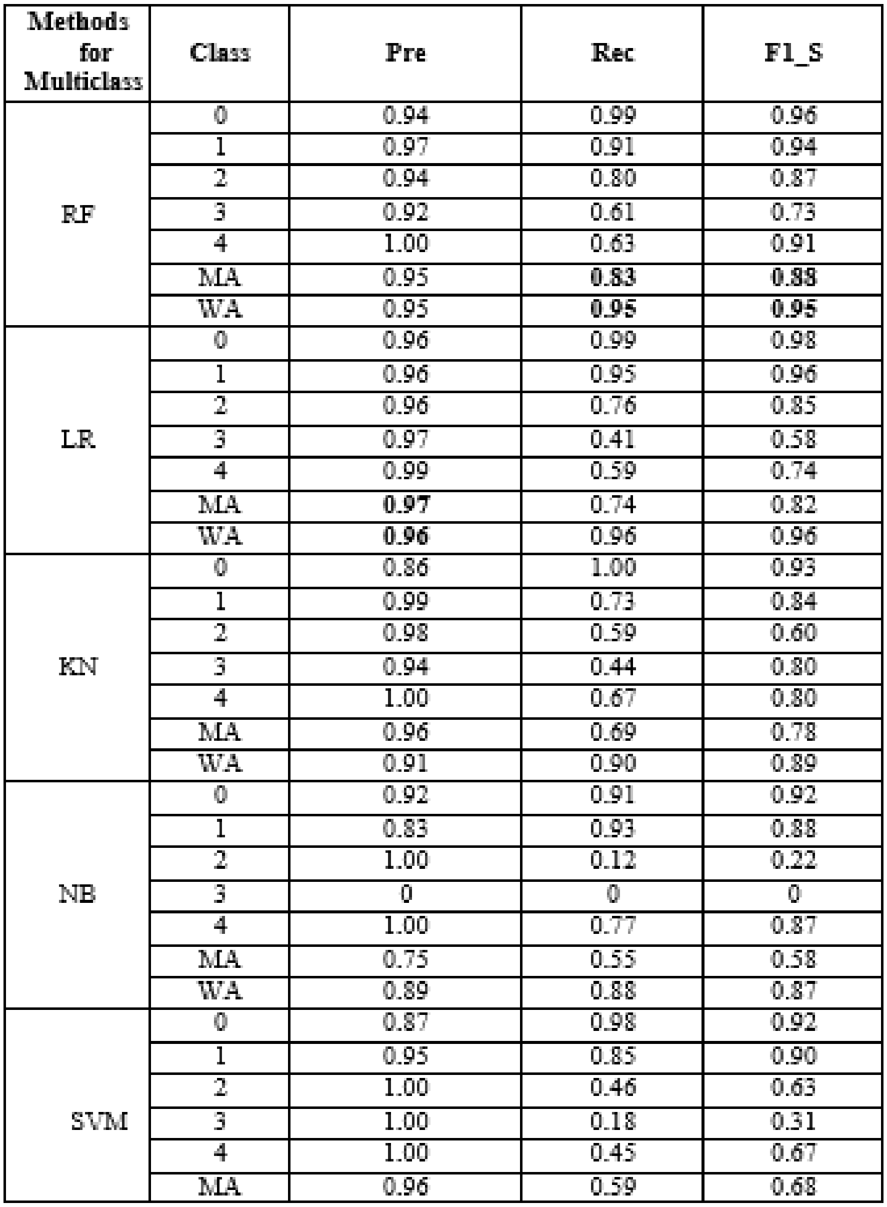
- **Logistic Regression (LR):** A linear model that estimates the probability of a binary outcome.

- **Random Forest (RF):** An ensemble method that combines multiple decision trees to improve classification accuracy and reduce overfitting.
- **Support Vector Machines (SVM):** A powerful classifier that constructs hyperplanes to separate different classes in the feature space.

Each classifier is assigned a weight based on its accuracy during training. The final prediction is determined by a weighted voting scheme, ensuring that the most reliable classifiers have a greater influence on the outcome.

### 4. EXPERIMENTAL RESULTS

The proposed models were tested on several real-world healthcare datasets, including the Pima Indians Diabetes Dataset, known for its use in diabetes research. The datasets were preprocessed to remove missing values and standardize feature scales before being subjected to feature selection using the CEFSM.



| Methods for Multiclass | Class | Pre | Rec | F1_S |
|---|---|---|---|---|
| RF | 0 | 0.94 | 0.99 | 0.96 |
| | 1 | 0.97 | 0.91 | 0.94 |
| | 2 | 0.94 | 0.80 | 0.87 |
| | 3 | 0.92 | 0.61 | 0.73 |
| | 4 | 1.00 | 0.63 | 0.91 |
| | MA | 0.95 | **0.83** | **0.88** |
| | WA | 0.95 | **0.95** | **0.95** |
| LR | 0 | 0.96 | 0.99 | 0.98 |
| | 1 | 0.96 | 0.95 | 0.96 |
| | 2 | 0.96 | 0.76 | 0.85 |
| | 3 | 0.97 | 0.41 | 0.58 |
| | 4 | 0.99 | 0.59 | 0.74 |
| | MA | **0.97** | 0.74 | 0.82 |
| | WA | **0.96** | 0.96 | 0.96 |
| KN | 0 | 0.86 | 1.00 | 0.93 |
| | 1 | 0.99 | 0.73 | 0.84 |
| | 2 | 0.98 | 0.59 | 0.60 |
| | 3 | 0.94 | 0.44 | 0.80 |
| | 4 | 1.00 | 0.67 | 0.80 |
| | MA | 0.96 | 0.69 | 0.78 |
| | WA | 0.91 | 0.90 | 0.89 |
| NB | 0 | 0.92 | 0.91 | 0.92 |
| | 1 | 0.83 | 0.93 | 0.88 |
| | 2 | 1.00 | 0.12 | 0.22 |
| | 3 | 0 | 0 | 0 |
| | 4 | 1.00 | 0.77 | 0.87 |
| | MA | 0.75 | 0.55 | 0.58 |
| | WA | 0.89 | 0.88 | 0.87 |
| SVM | 0 | 0.87 | 0.98 | 0.92 |
| | 1 | 0.95 | 0.85 | 0.90 |
| | 2 | 1.00 | 0.46 | 0.63 |
| | 3 | 1.00 | 0.18 | 0.31 |
| | 4 | 1.00 | 0.45 | 0.67 |
| | MA | 0.96 | 0.59 | 0.68 |

**Table.1:** Methods of Multiclass and corresponding Datasheets

### 4.1 Performance Evaluation

The performance of the proposed models was evaluated using standard metrics, including accuracy, precision, recall, and F1-score. The results indicate that the CECMSDML outperformed traditional classification methods across all datasets. Notably, the ensemble model achieved an accuracy of 87% on the Pima Indians Diabetes Dataset, significantly higher than the best-performing individual classifier.

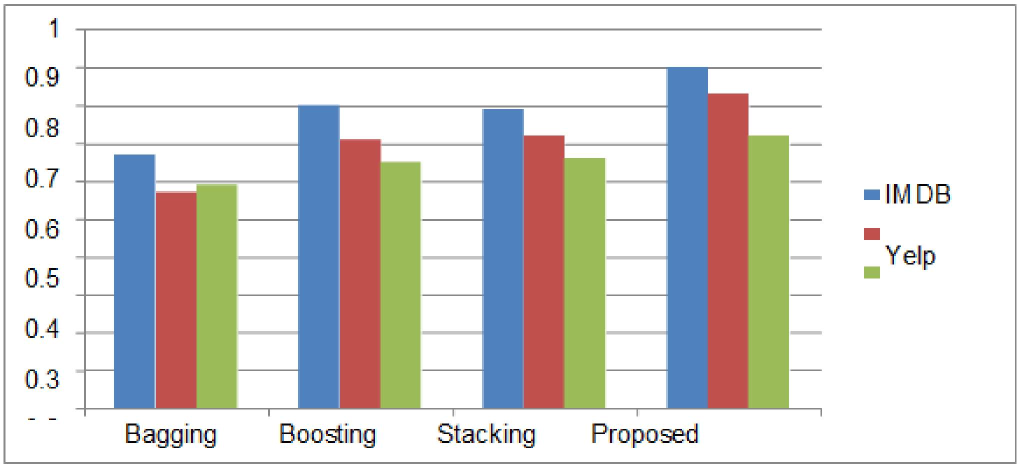


**Fig.2:** Benchmark Dataset Comparison of Existing and Proposed Work

## 5. SPECIFIC OUTCOME

The results of this study highlight the effectiveness of ensemble learning methods in the classification of T2D. By combining the strengths of multiple classifiers and focusing on the most relevant features, the proposed models were able to achieve superior performance compared to traditional methods. The study also underscores the importance of feature selection in managing high-dimensional healthcare data, which can significantly impact the accuracy of classification models. One of the key advantages of ensemble learning is its ability to improve the generalization of models, making them more robust to variations in the data. This is particularly important in healthcare, where patient data can be highly variable. The findings of this study suggest that ensemble learning methods could be a valuable tool in the early diagnosis and management of T2D, potentially leading to better patient outcomes.

## CONCLUSION

This paper presents a novel ensemble learning framework for the classification of T2D, demonstrating its potential to improve diagnostic accuracy. The proposed models, CEFSM and CECMSDML, effectively address the challenges of high-dimensional data and model selection, offering a robust approach to T2D classification. Future research could explore the application of these methods to other chronic diseases, further validating their utility in clinical practice. Additionally, integrating deep learning techniques into the ensemble framework could offer further improvements in performance.

## REFERENCES

1. Chou, S. M., Lee, T. S., Shao, Y. E., & Chen, I. F. (2004). Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications, 27(1), 133-142.

2. Kaur, G., & Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics, 14(1), 37-47.

3. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and Structural Biotechnology Journal, 15, 104-116.

4. Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., & Feng, D. (2015). Early diagnosis of Alzheimer's disease with deep learning. IEEE 12th International Symposium on Biomedical Imaging (ISBI), 1015-1018.

5. Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. PLoS Computational Biology, 15(6), e1006907.

6. Özçift, A. (2011). Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis. Computers in Biology and Medicine, 41(5), 265-271.

7. Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2021). Predicting Type 2 diabetes using hybrid machine learning approaches: A comprehensive review. Computers in Biology and Medicine, 134, 104588.

8. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

9. Priyanga, V., & Rathipriya, R. (2020). Prediction of type 2 diabetes using deep learning neural network: A review. Journal of Advanced Research in Dynamical and Control Systems, 12(7), 1419-1425.

10. Rai, H., & Chatterjee, K. (2020). Real-time big data analytics for prediction of heart disease using machine learning. Future Generation Computer Systems, 110, 785-797.

11. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533-536.

12. Saeed, F., DiCicco, L., Rashid, M., & Bassam, R. (2021). Ensemble learning methods for machine learning algorithms to predict Type 2 Diabetes. 2021 IEEE International Conference on Electro Information Technology (EIT), 1-5.

13. Samant, S., & Agarwal, P. (2018). Machine learning techniques for medical diagnosis. Journal of Biomedical Informatics, 90, 103-115.

14. Sun, J., & Reddy, C. K. (2013). Big data analytics for healthcare. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1525-1525.

15. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320.