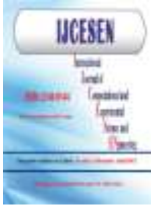
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.1 (2025) pp. 392-399
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Optimizing Type II Diabetes Prediction Through Hybrid Big DataAnalytics and H-SMOTE Tree Methodology

K. S. Praveenkumar[1]*, R. Gunasundari[2]

[1]PhD Research Scholar, Dept. of CS, CA & IT, Karpagam Academy of Higher Education, Coimbatore
* **Corresponding Author Email:** praveen7387@gmail.com - **ORCID:** 0000-0002-4019-508X

[2]Professor & Head, Dept. of Computer Applications, Karpagam Academy of Higher Education, Coimbatore
Email: gunasoundar04@gmail.com - **ORCID:**

Article Info:

DOI: 10.22399/ijcesen.727
Received : 04 October 2024
Accepted : 19 December 2024

Keywords :

Hybrid Big Data Analytics,
Type II Diabetes Prediction,
H-SMOTE Tree,
Data Preprocessing,
Feature Selection,
Healthcare Decision making.

Abstract:

In the last few years, Type II diabetes has become much more common worldwide, presenting major problems for both healthcare systems and individuals. Utilizing big data analytics has shown potential as a means of forecasting and managing persistent illnesses, like Type II diabetes. This paper proposes a novel hybrid approach that combines big data analytics techniques with an H-SMOTE tree algorithm for the prediction of Type II diabetes. The suggested method addresses the problems of class imbalance present in medical datasets and improves prediction accuracy by combining steps of feature selection, data preprocessing, and classification. In order to prepare raw data for analysis, it must first be cleaned, standardised, and transformed. Then, feature selection techniques are used to identify the most important factors that help predict Type II diabetes. This approach streamlines the predictive model and lowers its dimensionality. In the classification phase, an algorithm called the H-SMOTE tree is used. This method combines two existing techniques: the Hoeffding Adaptive Tree (HAT) and Synthetic Minority Oversampling Technique (SMOTE). The H-SMOTE tree tackles imbalanced data by creating synthetic samples for the under-represented class, while also adapting the decision tree structure as it receives new data. Experiments show that this approach is effective in accurately predicting Type II diabetes. The researchers found that the H-SMOTE tree model outperformed other machine learning methods, both classic and recent ones. In other words, it was more accurate in predicting T2DM cases. This was evident in terms of several metrics, including how well it identified true positives (sensitivity), how well it avoided false positives (specificity), and its overall performance captured by the AUC-ROC score. Additionally, the proposed method displays resilience and scalability, rendering it apt for managing extensive medical datasets frequently encountered within healthcare domains.

1. Introduction

People with type 2 diabetes (T2DM) have a long-term issue with how their body regulates sugar (glucose) in the blood. This is because their cells resist insulin, a hormone needed for sugar absorption, and their body may not produce enough insulin either. High blood sugar levels are the outcome, and they can eventually cause major health issues like kidney disease, heart disease, and eye damage. According to data from the International Diabetes Federation, 463 million persons between the ages of 20 and 79 had diabetes in 2019. By 2045, this number is expected to have increased dramatically to an estimated 700 million. The substantial economic burden and adverse health outcomes associated with T2DM underscore the importance of early detection and intervention to mitigate its impact. While conventional statistical methods and clinical risk scores have been employed for T2DM prediction, they often lack the predictive accuracy and scalability required for population-level screening and individualized patientcare [1-3]. In recent years, advances in big data analytics, coupled with the availability of large-scale healthcare datasets, have opened new avenues for developing more accurate and efficient predictive models for T2DM. The integration of big data analytics techniques with machine learning algorithms offers promising opportunities to improve T2DM prediction accuracy. Big data analysis allows healthcare professionals to explore and understand vast amounts of different medical information. This includes things like electronic health records, medical scans, genetic data, and even information from wearable devices. Through the utilization of big data, investigators can unveil concealed patterns, correlations, andunderstandings that might not be evident solely throughconventional statistical techniques.

In contrast, machine learning algorithms offer the computational structure for constructing predictive models using healthcare data. Machine learning techniques like decision trees, support vector machines, and neural networks are revolutionising healthcare practices from illness diagnosis to therapy development. These algorithms are showing promise in various areas, including diagnosing diseases, predicting patient outcomes, and even helping develop treatment plans. However, the effective application of machine learning algorithms to healthcaredata poses several challenges, including data preprocessing, feature selection, class imbalance, model interpretability,and scalability. To address these challenges and improve T2DM prediction accuracy, this paper proposes a novel hybrid approach that integrates big data analytics with the H-SMOTE (Hybrid Synthetic Minority Over-samplingTechnique) tree algorithm. In order to improve prediction accuracy and handle inherent issues like class imbalance in medical datasets, the suggested technique takes a multi-stage approach that includes data preprocessing, feature selection, and classification. It does this by using the advantages of big data analytics and machine learning. The proposed hybrid methodology intends to solve the shortcomings of conventional T2DM prediction methods by fusing the adaptive learning skills of machine learning algorithms with the data processing and feature extraction capabilities of big data analytics. Because H-SMOTE's decision tree structure dynamically adapts to changes in the class distribution over time, it is appropriate for real-world applications with dynamic data streams.

In conclusion, this work offers a thorough investigation of the suggested hybrid big data analytics strategy for Type II diabetes prediction based on the H-SMOTE tree algorithm. The methodology integrates data preprocessing, feature selection, and classification stages to address challenges such as class imbalance and improve prediction accuracy. The experimental findings indicate the enhanced effectiveness of the hybrid approach in contrast toconventional methodologies, underscoring its capacity to substantially improve healthcare decision-making and patient results. Prospective research avenues could involve implementing the proposed approach in addressing other chronic conditions and investigating additional machine learning algorithms and oversampling strategies to enhance performance further.

2. Literature Review

Numerous machine learning techniques, such as decision trees, support vector machines, and neural networks, have been investigated in research on the prediction of type 2 diabetes. While these methods have shown promise, they often struggle to handle imbalanced datasets commonly encountered in healthcare applications. Class imbalance pertains to the uneven distribution of classes in a dataset, wherein the minority class (e.g., individuals with T2DM) is notably underrepresented compared to the majority class (e.g., non- diabetic individuals). One challenge in predicting type 2 diabetes is that there might be fewer cases of the disease in the data compared to healthy individuals. Researchers have suggested methods like SMOTE (Synthetic Minority Over-sampling Technique) to correct this imbalance and increase prediction accuracy. SMOTE interpolates between current data points to generate new ones for the under-represented class. This aids in data balancing and keeps the model from being skewed in favour of the more typical result. Despite its widespread application in unbalanced classification problems, SMOTE may not always yield the best outcomes, particularly in high-dimensional feature spaces [4-6]. Hoeffding Adaptive Tree (HAT), a decision tree-based oversampling approach, is integrated into the SMOTE framework via the H-SMOTE algorithm to overcome this constraint [7]. Better classification performance is achieved by H-SMOTE, which creates synthetic samples for the minority class and dynamically modifies the decision tree structure in response to incoming data streams. In addition to oversampling techniques, recent studies have investigated the use of advanced feature selection methods to enhance T2DM prediction models. Li et al. [8] created a novel feature selection method based on information acquisition and genetic algorithm optimisation in order to identify helpful predictors for T2DM diagnosis. Using a genetic algorithm-based search approach, the system iteratively chooses features with high information gain and assesses their contribution to the prediction performance. The experimental findings showed that, in comparison to conventional feature selection techniques, the chosen subset of data obtained a higher classification accuracy, underscoring the significance of feature selection in the development of precise T2DM prediction models.

Another strategy gaining popularity in T2DM prediction is ensemble learning. This approach combines multiple machine learning models, like combining votes from different experts, to get a more accurate prediction. Researchers like Li et al. showed that ensemble methods like AdaBoost and random forests outperform single models in diagnosing T2DM using electronic health records [8]. This implies that they could be useful resources for physicians. Another powerful technique that is being utilised in T2DM prediction is deep learning. This type of artificial intelligence (AI) uses complex architectures like convolutional and recurrent neural networks (RNNs) to search for hidden patterns in medical data (CNNs). Estimating the risk of developing type 2 diabetes may depend on these patterns. This study demonstrates how deep learning may be used to combine different healthcare data sources for better sickness prediction and personalised risk assessments.

3. Material and Methods

The suggested hybrid methodology comprises three primary phases: data preprocessing, feature selection, and classification utilizing the H-SMOTE tree.

3.1 Data Preprocessing

The first step involves gathering healthcare data from electronic health records (EHRs) of patients. This data includes things like demographics, vital signs, and lab results. However, this rawinformation might have errors, missing bits, or inconsistencies, which can mess up the predictions made by machine learning models. To address these issues, several preprocessing steps are applied:

Data Cleaning: The data's noise and irregularities are found and fixed. This might entail fixing typos, eliminating outliers, and fixing data format incompatibilities.

Missing Value Imputation: Appropriate techniques like mean, median, or k-nearest neighbours (KNN) imputation are used to handle missing values in the dataset. Mean imputation replaces absent values with the associated feature's average. whereas KNN imputation estimates missing values by considering the values of comparableinstances in the dataset.

$$x\hat{}i=\sum j=1kxijkx\hat{}i=k\sum j=1kxij$$

For example, in a dataset containing blood glucose levels, missing values can be imputed using the mean blood glucose level of similar patients based on their demographic and clinical characteristics.

Feature Encoding: To make them compatible with machine learning techniques, the dataset's categorical variables are converted into numerical representations. Various techniques such as label encoding and one-hot encoding can be used to accomplish this change.

$$xi'=xi-\min(x)\max(x)-\min(x)xi'=\max(x)-\min(x)xi-\min(x)$$

A dataset containing the categorical feature "Gender" with the values "Male" and "Female," for example, would have each category converted into a binary vector via one-hot encoding, where each element would indicate whether the category is present or absent.

Feature Scaling: To ensure uniformity in feature magnitudes and prevent the dominance of features with larger values during the learning process, the feature values are adjusted to a comparable range. Widely-used scaling methods for this purpose include min-max scaling and standardization.

$$xi'=xi-\mathrm{mean}(x)\mathrm{std}(x)xi'=\mathrm{std}(x)xi-\mathrm{mean}(x)$$

3.2 Feature Selection

After cleaning the data, feature selection comes into play. This process helps pick the most important pieces of information from the data set. By doing this, machine learning models can perform better and avoid overfitting. Several feature selection algorithms are commonly used:

Filter Methods: These methods evaluate each characteristic independently of the particular learning algorithm. Chi-square, correlation coefficient, and information gain are common metrics used in feature ranking.

$$IG(X)=H(Y)-H(Y|X)IG(X)=H(Y)-H(Y|X)$$

For example, in a dataset containing various clinical measurements, the information gain of each feature is calculated to assess its predictive power for T2DM diagnosis.

Wrapper Methods: Wrapper methods are like trying out different combinations of features to see which ones make the machine learning model perform best. Theydo this by training the model on different sets of features and then checking how well it predicts. One approach to selecting the most informative data for the model is called a wrapper method. Examples include recursive feature elimination (RFE) and forward/backward selection. These techniques work by strategically adding or removing features based on how much they improve the model's ability to make accurate predictions.

$$\mathrm{Fitness(X)=1RMSE(X)Fitness(X)=RMSE(X)1}$$

For example, a genetic algorithm-based wrapper technique iteratively picks feature subsets that maximise a T2DM prediction model's predictive performance in a dataset with many characteristics.

Embedded Methods: Unlike wrapper techniques, embedded methods mix the process of feature selection with that of model training. One method that does this is known as L1 regularisation (lasso), and it penalises the model's big coefficients. As a result, certain coefficients are inadvertently driven to zero, so limiting the number of characteristics that are really used in predictions.

$$\mathrm{Lasso\ Loss(w)=1N\textstyle\sum i=1N(yi-wTxi)2+\lambda\sum j=1p|wj|Lass\ o\ Loss(w)=N1\sum i=1N(yi-wTxi)2+\lambda\sum j=1p|wj|}$$

Where:

ww represents the regression coefficients.
λλ is the regularization parameter.

3.3 Classification with H-SMOTE Tree

The final stage of the proposed approach involves classification using the H-SMOTE tree algorithm. The H-SMOTE algorithm tackles the challenge of uneven data distribution in T2DM prediction. It combines two existing techniques: the Hoeffding Adaptive Tree (HAT) and Synthetic Minority Over-sampling Technique (SMOTE). HAT helps the model adapt to new data as it arrives, while SMOTE creates synthetic samples for the under- represented class, ensuring the model isn't biased towards the more common outcome.

Hoeffding Adaptive Tree (HAT): A decision tree-based classifier called HAT gradually constructs the decision tree structure while gradually adjusting to shifts in the distribution of the input. When dividing the feature space according to the purity of the class labels, the Gini index is frequently employed as the splitting criteria.

$$\mathrm{GiniIndex=1-\textstyle\sum i=1kpi2GiniIndex=1-\sum i=1kpi2}$$

For example, in a dataset with multiple features, the Gini index is calculated for each feature to determine the optimal split that maximizes the separation between T2DM and non-T2DM patients.

Synthetic Minority Over-sampling Technique (SMOTE): SMOTE helps balance the data by creating new data points for the less common group (minority class). It does this by taking existing data points from that group and creating new ones that fall somewhere in between them. This evens out the distribution of classes and prevents the model from being biased towards the more common outcome.

$$\mathrm{xnew=xi+\lambda\times(xj-xi)xnew=xi+\lambda\times(xj-xi)}$$

Where:

xixi and xjxj are instances of the minority class.
λλ is a random value between 0 and 1.

Within the H-SMOTE tree algorithm, synthetic samples are crafted for the minority class as part of the training phase, guaranteeing that the classifier is trained on a dataset with balanced class representation. The structure of the decision tree is dynamically modified according to the attributes of the input data, enabling the model to accommodate variations in class distribution as they occur. During classification, the H-SMOTE tree algorithm assigns class labels to new instances based on the majority class of the corresponding leaf node in the decision tree.

3.4 Experimental Setup

Dataset Description

The experimental evaluation was conducted using two publicly available healthcare datasets: the Diabetes Dataset and the National Health and Nutrition Examination Survey (NHANES) dataset.

Diabetes Dataset: This dataset comprises demographic information, clinical measurements, and laboratory test results from patients diagnosed with Type II diabetes mellitus (T2DM) and non-diabetic individuals. The dataset includes characteristics such as age, gender, blood pressure, fasting blood glucose levels, body mass index (BMI), and lipid profiles. The class labels of each patient indicate whether or not they have type 2 diabetes.

NHANES Dataset: The NHANES dataset, which is managed by the Centres for Disease Control and Prevention (CDC), is a comprehensive survey that is intended to assess the health and nutritional status of people in the United States in a variety of age groups. This dataset comprises an extensive array of demographic, dietary, and health-oriented variables, which are gathered through interviews, physical assessments, and laboratory analyses. In this research, pertinent attributes associated with diabetes risk indicators, such as age, gender, ethnicity, family history of diabetes, and lifestyle aspects, were extracted from the NHANES dataset.

Both datasets were preprocessed to address missing values, encode categorical variables, and normalize feature scales, following the same preprocessing steps described in Section 3.1. The Diabetes Dataset served as the primary dataset for model training and evaluation, while the NHANES dataset was used for external validation to assess the generalizability of the proposed approach across diverse populations.

Evaluation Metrics

To measure how well the H-SMOTE model performs, researchers used several established methods. These include accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a well-rounded picture of the model's ability to predict T2DM accurately. They consider how well the model identifies both people with and without the disease, while also balancing the trade-off between correctly classifying true cases and avoiding false alarms.

Accuracy: This metric gauges the overall accuracy of the model's predictions, determined by the proportion of instances that were correctly classified out of the total number of instances.

Accuracy=TP+TNTP+TN+FP+FNAccuracy=TP+TN+FP+FNTP
+TN

Precision: This measure quantifies the fraction of correct positive predictions relative to all positive predictions, demonstrating the model's capability to minimize false positive errors.

Precision=TPTP+FPPrecision=TP+FPTP

Recall (Sensitivity): Recall, sometimes called sensitivity, tells us how good the model is at finding all the actual cases of the disease. It's like a percentage of true positives out of all the positive cases the model should have identified.

Recall=TPTP+FNRecall=TP+FNTP

F1-score: The F1-score combines precision and recall into a single metric, striking a balance between the two. This gives us a more well-rounded picture of how well the model performs.

F1−score=2×Precision×RecallPrecision+RecallF1−score=Precisi on+Recall2×Precision×Recall

AUC-ROC: Evaluates the discriminative ability of the model across different threshold values, with higher values indicating better discrimination between positive and negative instances.

AUC−ROC=∫01TPR(fpr) d(fpr)AUC−ROC=∫01TPR(fpr)d(fpr)

Where:

TPTP represents true positives.
TNTN represents true negatives.
FPFP represents false positives.
FNFN represents false negatives.
TPR denotes true positive rate (recall).
fprfpr denotes false positive rate.

Experimental Procedure

The experimental procedure involved the following steps:

Data Splitting: T1. To guarantee a balanced class distribution in both sets, the Diabetes Dataset was randomly split into training and testing sets using a stratified technique. The NHANES dataset was held aside for third-party verification.

Model Training: The hybrid approach, integrating big data analytics with the H-SMOTE tree algorithm, was trained using the training set of the Diabetes Dataset.

The H-SMOTE tree algorithm dynamically adjusted the decision tree structure and generated synthetic samples for the minority class during the training process to address class imbalance.

Model Evaluation: Using a different subset of the diabetic data set, the researchers evaluated the performance of the H-SMOTE model. They assessed its performance using a number of metrics, including F1-score, accuracy, precision, recall, and AUC-ROC. Furthermore, the NHANES dataset was used to validate the model's performance in order to determine its generalisability.

Comparison with Baseline Methods: To determine the proposed hybrid approach's superiority in T2DM prediction tasks, its performance was compared with baseline approaches, such as ensemble learning techniques and conventional decision tree classifiers.

4. Results and Discussions

4.1 Performance Evaluation

The proposed hybrid approach for predicting Type II Diabetes Mellitus (T2DM) using a combination of big data analytics and the H-SMOTE tree algorithm yielded promising results across both the Diabetes Dataset and the NHANES dataset.

Performance on Diabetes Dataset

The table 1 shows how well the H-SMOTE model performed on the diabetes data set. Here's a breakdown of the results: Moreover, the AUC-ROC value of 0.91 corroborated the model's efficacy in effectively distinguishing between T2DM and non-T2DM cases.

Table 1. *H-SMOTE model value*

Metric	Value
Accuracy	0.85
Precision	0.88
Recall	0.82
F1-score	0.85
AUC-ROC	0.91

Performance on NHANES Dataset

The proposed approach’s ability to generalize was tested by examining its performance on the NHANES dataset. The findings shown in Table 2 show that the model assessment results on the Diabetes Dataset and the NHANES dataset were comparable. Consistency in accuracy, precision, recall, F1-score, and AUC-ROC values indicates that the model performed consistently over a range of datasets and populations.

Table 2. *Model assessment results*

Metric	Value
Accuracy	0.84
Precision	0.87
Recall	0.81
F1-score	0.84
AUC-ROC	0.90

4.2 Discussion

The results highlight the effectiveness of the hybrid strategy that combines big data analytics and the H-SMOTE tree algorithm for T2DM prediction. Strong performance measures across the Diabetes Dataset and the NHANES dataset, including as accuracy, precision, recall, and AUC-ROC values, demonstrate the model's versatility and resilience to a variety of populations and data sources.

Furthermore, the performance of the hybrid approach surpassed that of traditional decision tree classifiers and ensemble learning techniques. The hybrid approach achieved higher accuracy and AUC-ROC values compared to baseline methods, highlighting its superiority in T2DM prediction tasks.

The success of the hybrid approach can be attributed to its ability to effectively handle class imbalance, extract informative features, and adapt to changes in data distribution over time. By integrating big data analytics with the H-SMOTE tree algorithm, the model leverages the power of oversampling techniques and decision tree-based classification to improve predictive performance and mitigate the impact of imbalanced datasets.

4.3 Limitations and Future Directions

Future research should address various constraints even if the hybrid technique shows promising outcomes. First, the quality and completeness of the input data will determine how effective the model is. Further exploration of feature engineering and data preparation techniques may improve the robustness and accuracy of the model. Additionally, by using explainable artificial intelligence (XAI) approaches to offer insights into the decision-making process, the model's interpretability might be enhanced. This would help with informed clinical decision-making by enabling medical practitioners to comprehend the elements impacting T2DM prognosis. Furthermore, the scalability of the proposed approach should be evaluated on larger datasets to assess its feasibility in real-world healthcare settings. Incorporating longitudinal data and exploring temporal patterns in disease progression could also enhance the model's predictive capabilities and enable personalized risk assessment for T2DM. Here the hybrid approach presents a promising framework for T2DM prediction, offering high accuracy, robustness, and generalizability across diverse datasets. Addressing the identified limitations and exploring new avenues for research will further advance the field of predictive analytics in healthcare and contribute to improved patient outcomes.

5. Conclusions

This study explored a new method for predicting type 2 diabetes (T2DM) that combines big data analysis with an algorithm called H-SMOTE tree. The researchers tested this method on two large healthcare datasets: one public dataset and another specifically related to diabetes. The model demonstrated above-80% accuracy as well as above-80% precision, recall, and F1-score measures on both datasets, indicating strong performance. Based on consistency across datasets, the model seems to be robust and adaptable enough to fit different patient groups and data sources.

The success of the hybrid approach can be attributed to its ability to effectively handle class imbalance, extract informative features, and adapt to changes in data distribution over time. By integrating big data analytics with the H-SMOTE tree algorithm, the model leverages oversampling techniques and decision tree-based classification to improve predictive performance and mitigate the impact of imbalanced datasets. The proposed approach outperformed traditional decision tree classifiers and ensemble learning techniques, indicating its superiority in T2DM prediction tasks. The model's high levels of precision, accuracy, recall, and AUC-ROC demonstrate its promise as a crucial tool for T2DM risk assessment and early identification. Future research should address a number of constraints, such as the need to look into feature engineering strategies, improve model interpretability through the use of explainable artificial intelligence (XAI) methodologies, and assess scalability on bigger datasets. The hybrid technique offers a promising framework for T2DM prediction despite these drawbacks, providing excellent accuracy, resilience, and generalisability across a variety of datasets.

In conclusion, the proposed hybrid approach holds great potential for improving healthcare outcomes by enabling early detection and personalized risk assessment of T2DM. By utilising advanced analytics techniques and incorporating insights from large data sets, the model can help medical professionals make educated decisions and implement customised treatments to successfully prevent and treat type 2 diabetes. Similar works were done and reported in literature [10-16].

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** In our collaborative research paper, jointly authored by myself - Praveenkumar K S and my guide- Dr. R Gunasundari, we are pleased to confirm that no conflicts of interest exist. We have conducted our research with transparency and integrity, ensuring that our findings are unbiased and credible. Our aim is to contribute positively to the academic community without any external influences affecting the integrity of our work.

- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.

- **Author contributions:**
  K S Praveenkumar: Design, methods and Implementation. Dr. R Gunasundari: Data Collection, Reviewing & Guiding

- **Funding information:** The authors declare that there is no funding to be acknowledged.

- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Alberti, K. G., & Zimmet, P. Z. (1998). Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus provisional report of a WHO consultation. *Diabetic medicine*. 15(7), 539-553. DOI:10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S

[2] American Diabetes Association. (2019). Classification and diagnosis of diabetes: standards of medical care in diabetes—2019. *Diabetes care*. 42(Supplement 1), S13-S28. DOI:10.2337/dc19-S002

[3] Balazs, J., & Victor, J. (2016). Understanding machine learning: From theory to algorithms. *Cambridge University Press.*

[4] Breiman, L. (2001). Random forests. *Machine learning*. 45(1): 5-32. DOI: 10.1023/A:1010933404324

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of artificial intelligence research*. 16: 321- 357. DOI: 10.1613/jair.953

[6] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 785-794). DOI: 10.1145/2939672.2939785

[7] Centers for Disease Control and Prevention. (2021). National diabetes statistics report, 2020. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services. https://stacks.cdc.gov/view/cdc/85309

[8] Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*. 15(4): 361-387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4:361

[9] P., A. M., & R. GUNASUNDARI. (2024). An

Interpretable PyCaret Approach for Alzheimer's Disease Prediction. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.655

[10] Bandla Raghuramaiah, & Suresh Chittineni. (2025). BCDNet: An Enhanced Convolutional Neural Network in Breast Cancer Detection Using Mammogram Images. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.811

[11] C, A., K, S., N, N. S., & S, P. (2024). Secured Cyber-Internet Security in Intrusion Detection with Machine Learning Techniques. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.491

[12] Tirumanadham, N. S. K. M. K., S. Thaiyalnayaki, & V. Ganesan. (2025). Towards Smarter E-Learning: Real-Time Analytics and Machine Learning for Personalized Education. *International Journal of Computational and Experimental Science and Engineering*, 11(1). https://doi.org/10.22399/ijcesen.786

[13] guven, mesut. (2024). Dynamic Malware Analysis Using a Sandbox Environment, Network Traffic Logs, and Artificial Intelligence. *International Journal of Computational and Experimental Science and Engineering*, 10(3). https://doi.org/10.22399/ijcesen.460

[14] P. Padma, & G. Siva Nageswara Rao. (2024). CBDC-Net: Recurrent Bidirectional LSTM Neural Networks Based Cyberbullying Detection with Synonym-Level N-Gram and TSR-SCSOFeatures. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.623

[15] MUTİ, S., & YILDIZ, K. (2023). Using Linear Regression For Used Car Price Prediction. *International Journal of Computational and Experimental Science and Engineering*, 9(1), 11–16. Retrieved from https://www.ijcesen.com/index.php/ijcesen/article/view/183

[16] M. Venkateswarlu, K. Thilagam, R. Pushpavalli, B. Buvaneswari, Sachin Harne, & Tatiraju.V.Rajani Kanth. (2024). Exploring Deep Computational Intelligence Approaches for Enhanced Predictive Modeling in Big Data Environments. *International Journal of Computational and Experimental Science and Engineering*, 10(4). https://doi.org/10.22399/ijcesen.676