IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Predicting Diabetes Risk In Pima Indian Women Using Deep Learning Model

Palak Majevadiya[1]

M.E. Student[1]
Department of Computer Engineering
KIRC, KALOL
E-mail:

Prof .Nidhi Joshi [2]

Assistant Professor[2]
Department of Computer Engineering
KIRC, KALOL
E-mail:

Abstract-Diabetes mellitus remains a pervasive public health issue, particularly among certain ethnic groups such as the Pima Indian community. This study leverages deep learning methodologies to forecast the risk of Type 2 diabetes in Pima Indian women using the well-established Pima Indian Diabetes dataset. Employing Artificial Neural Networks (ANNs), including multilayer perception architectures, the model identifies intricate, nonlinear patterns in clinical data such as glucose levels, BMI, and age. Results indicate that the deep learning model achieves superior accuracy and robustness compared to traditional machine learning methods, underscoring its potential as a viable tool for early detection and personalized intervention strategies in diabetes.

I.INTRODUCTION

Diabetes mellitus is a chronic metabolic disorder marked by elevated blood glucose levels due to inadequate insulin production or resistance. Type 2 diabetes, the most prevalent form, disproportionately affects specific populations, including Pima Indian women. Contributing factors range from genetic predisposition and obesity to socio-economic limitations and poor access to healthcare. Early prediction and management of diabetes are crucial for mitigating long-term complications such as cardiovascular disease, kidney failure, and neuropathy.

Given the advent of machine learning and deep learning technologies, healthcare has witnessed transformative advancements. These models can process large volumes of structured and unstructured data to detect patterns invisible to traditional methods. This research investigates the efficacy of deep learning models in predicting diabetes risk, offering a scalable, accurate, and

non-inv

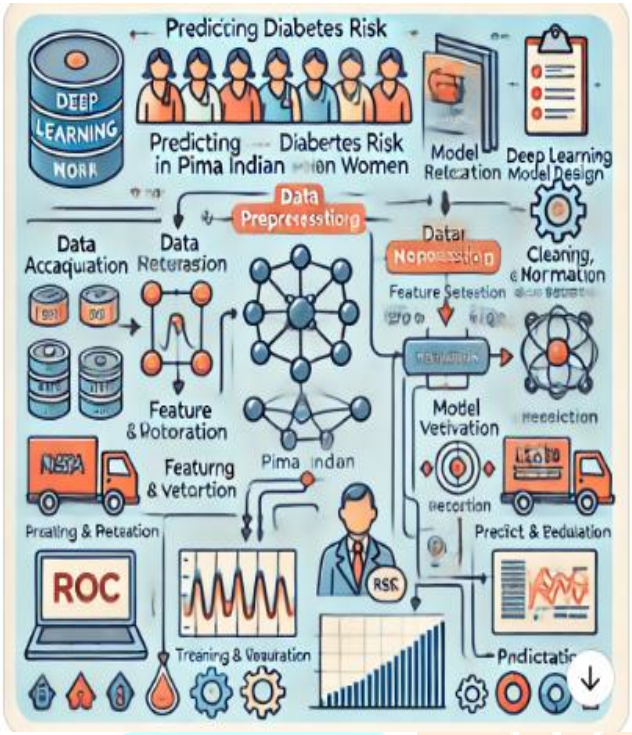

Fig 1.Schematic diagram of a Diabetes system

II. BACKGROUND THEORY

Previous studies have explored various machine learning classifiers—such as logistic regression, decision trees, support vector machines, and ensemble models—for diabetes prediction. However, deep learning models, particularly ANNs and CNNs, have demonstrated superior performance due to their capacity to model complex, high-dimensional relationships.

Literature reveals that hybrid approaches combining feature selection and oversampling techniques improve accuracy. Models such as Random Forests, Gradient Boosting, and Deep Neural Networks have been tested on the Pima Indian dataset, achieving varying degrees of precision, recall, and F1 scores. Despite these efforts, challenges like dataset imbalance, limited diversity, and interpretability persist.

III. RELATED WORK

[1]. Title: Diabetes detection based on machine learning and deep learning approaches Multimedia Tools and Applications (Springer)-2023

Purpose: This study aims to explore and analyze the potential of ML and DL techniques in creating accurate, efficient, and cost-effective diabetes detection models. Moreover, to address the limitations of traditional lab-based diabetes detection methods, which are invasive, expensive, and time-consuming? Key idea: • The study emphasizes the importance of high-quality datasets and addresses common issues like missing data and dimensionality reduction. • Proposes hybrid models combining ML and DL algorithms to improve accuracy and robustness in diabetes detection. • Investigates the impact of feature selection and oversampling techniques on the performance of diabetes detection models. • Highlights the growing demand for diabetes detection tools using non-invasive measurements. Results: The study demonstrated the feasibility of using non-invasive datasets for diabetes detection, with promising but less reliable outcomes than invasive methods. The deep learning models, especially convolution neural networks (CNNs) and deep neural networks (DNNs), generally outperform traditional ML models in terms of accuracy. In addition to that Feature selection techniques and data pre-processing improved the efficiency of the models, but their effectiveness varied based on dataset quality and size. Limitations: The lack of large-scale, high-quality, and diverse datasets is a significant barrier to achieving reliable and generalized models and the Over fitting was identified as a potential issue, especially in tree-based models like Random Forests when improperly tuned.

[2] Title: Machine Learning-Based Diabetes Classification and Prediction for Healthcare Applications [Hindawi-2022]

Purpose: The paper aims to enhance the diagnosis of diabetes using machine learning techniques. The primary objective is to classify diabetes cases accurately by comparing the performance of six machine learning classifiers. Key idea: The study evaluates the performance of various machine learning models, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), XGBoost, Random Forest, and Ad boost, to classify diabetes cases. Feature selection using Principal Component Analysis (PCA) was applied to optimize the model's performance by reducing noise and redundancy in the dataset. Result: Ad boost achieved the highest accuracy of 83%, followed closely by Random Forest and Logistic Regression.SVM had an accuracy of 82.46%, while Decision Tree and XGBoost performed below 80%. Ad boost also performed well across

other metrics like F1-score and precision.PCA was instrumental in selecting significant features and improving model performance. Limitations: The dataset was limited to female patients of Pima Indian heritage, restricting the generalizability of the findings to other populations. While PCA improved performance, further exploration of alternative feature selection techniques could be beneficial.

[3]Title: Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms [Neural Computing and Applications– 2023] Purpose: The paper aims to develop an e-diagnosis system for type 2 diabetes classification within an Internet of Medical Things (IoMT) environment using interpretable machine learning (ML) models. It seeks to address the lack of trust in ML systems caused by their "black-box" nature by employing interpretable models (Naïve Bayes, Random Forest, and J48 Decision Tree). This system is intended to provide early diagnosis support, which can improve health outcomes and reduce treatment costs, particularly in underserved regions. Key idea: The research evaluates three interpretable ML models—Naïve Bayes, Random Forest, and J48 Decision Tree—on the Pima Indian Diabetes dataset. The goal is to compare their performance across various metrics (accuracy, precision, sensitivity, specificity, F1-score, and AUC) to identify the best algorithm for binary diabetes classification. The study also incorporates feature selection techniques, including Principal Component Analysis (PCA) and k-means clustering, to improve interpretability and model performance. Result: Random Forest Achieved the highest overall accuracy (79.57%) on the full dataset with superior precision (89.40%), F1-score (85.17%), and AUC (86.24%), while Naïve Bayes Excelled with accuracy (79.13%) on the 3-feature subset, demonstrating the importance of fine-tuned feature selection. Limitations: The Pima Indian Diabetes dataset is relatively small and lacks diversity, which could limit generalizability. Features reduction method is **[4] Title: Prediction of diabetes disease using machine learning algorithms [IAES International Journal of Artificial Intelligence (IJ-AI)-2022]** Purpose: The paper aims to develop a predictive model for diagnosing diabetes using machine learning algorithms. The objective is to create an effective and efficient model with high accuracy, precision, and reduced processing time to enhance early detection and management of diabetes. Key idea: The study evaluates four machine learning algorithms—Logistic Regression (LR), KNearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting (GB)—on the Pima Indian Diabetes dataset. It compares their performance on metrics such as accuracy, precision, recall, and F1-score to identify the most effective model for predicting diabetes. Result: Gradient Boosting achieved the highest accuracy (81.25%) and outperformed others in precision (0.76), recall (0.63), and F1-score (0.70). Logistic Regression closely followed Gradient Boosting with an accuracy of 81%.SVM scored 80%, and KNN scored the lowest accuracy of 78%. Limitations: The models rely heavily on predefined features, and additional feature engineering might enhance results. It does not explore more advanced or hybrid machine learning techniques like deep learning or ensemble methods that may improve performance further.

[5]Title: A comparison of machine learning algorithms for diabetes prediction [Elsevier, 2021]

Purpose: The paper aims to compare different machine learning (ML) algorithms to enhance the prediction accuracy of diabetes using the Pima Indian Diabetes Dataset (PIDD). By identifying the most effective methods, it seeks to facilitate early diabetes detection for better disease management. Key idea: The study evaluates four machine learning algorithms—Logistic Regression (LR), KNearest Neighbors (KNN), Support Vector Machine (SVM), and Gradient Boosting (GB)—on the Pima Indian Diabetes dataset. It compares their performance on metrics such as accuracy, precision, recall, and F1-score to identify the most effective model for predicting diabetes. Result: Logistic Regression (LR) and Support Vector Machine (SVM) performed best among ML methods, achieving 78.85% accuracy with train/test splitting K-Nearest Neighbor (KNN) and Adaptive Boosting (AB) also showed high accuracy (79.42%) with the same method. Limitations: The findings are based solely on the Pima Indian Diabetes Dataset, limiting generalizability to other populations or datasets. While KNN is noted for reduced processing time, its accuracy lags behind other models, limiting its utility. Average of different attributes.

V. PROPOSED FRAMEWORK

1. Data Layer: Data Acquisition & Storage

- **Dataset**: Pima Indian Diabetes Dataset (UCI/Kaggle)
- **Features**: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome
- **Storage**: CSV/Excel-based dataset integrated into the framework via Pandas/Numpy

2. Data Preprocessing Layer

- **Missing Value Treatment**: Imputation with mean/median or KNN imputation
- **Normalization**: Min-Max scaling or Standardization
- **Outlier Detection**: IQR/Z-Score method
- **Feature Engineering**:

Feature selection using PCA or correlation matrix

Optional: Synthetic data balancing using SMOTE

3. Deep Learning Model Layer

Model Type: Artificial Neural Network (ANN)

Architecture:

Input Layer: 8 features

Hidden Layers: 2–3 dense layers with ReLU activation

Dropout Layer: For regularization (e.g., 0.2 dropout rate)

Output Layer: 1 Neuron with Sigmoid activation (for binary classification)

Optimizer: Adam

Loss Function: Binary Cross-Entropy

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC

4. Training & Validation Layer

Dataset Split: 70% training, 30% testing

Validation: K-Fold Cross Validation (optional)

Hyper parameter Tuning: Learning rate, epochs, batch size (using Grid Search or manual tuning)

Over fitting Control: Early stopping, dropout layers

5. Prediction & Output Layer

Output: Diabetes Risk Prediction (Positive / Negative)

Confidence Score: Probability output (0 to 1)

User Interface (Optional):

Dashboard for inputting new patient data

Risk assessment visualization

Exportable reports for clinical review

6. Interpretability & Deployment (Optional/Future Work)

Model Interpretation:

SHAP or LIME to explain model decisions

Deployment:

Flask web app or Stream lit interface

Integration with electronic health record (EHR) systems.

Data Collection: The collected data should be diverse and representative of different population groups to account for variations in health outcomes across age, gender, ethnicity, and geographic location. Additionally, the dataset must be large enough to enable effective training of the deep learning model, ensuring robust performance and reducing the risk of over fitting. By gathering a rich and diverse dataset, this project aims to

build a model that not only predicts diabetes risk with high accuracy but also provides actionable insights that can support personalized healthcare strategies for individuals at risk.

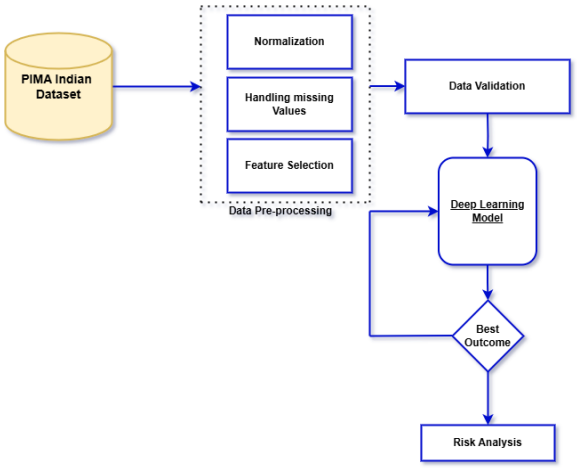

Fig 2: Proposed Framework

VII. Problem Statement and Objectives

ProblemStatement:
Traditional diagnostic tools for diabetes are invasive, time-consuming, and often inaccessible in under-resourced settings. There is a critical need for an automated, reliable, and early-stage risk prediction system that utilizes readily available health metrics.

Objectives:

- Preprocess and clean the Pima Indian dataset for optimal model input.
- Design a deep learning model capable of predicting diabetes risk with high accuracy.
- Evaluate performance using metrics such as accuracy, precision, recall, and AUC-ROC.
- Compare the performance with conventional machine learning algorithms.
- Enhance interpretability for integration into clinical workflows.

VIII. Methodology

1 Dataset Description

The dataset consists of clinical parameters from Pima Indian women aged 21 years and older. Features include:

- Number of pregnancies
- Glucose concentration
- Blood pressure
- Skin thickness
- Insulin level
- Body Mass Index (BMI)
- Diabetes pedigree function
- Age
- Diabetes outcome (binary)

2 Data Preprocessing

Data preprocessing involved:

- Handling missing values via imputation
- Normalizing numeric variables
- Encoding categorical features
- Removing outliers
- Applying feature selection techniques (e.g., PCA)

3 Model Architecture

A multilayer perceptron (MLP) ANN was implemented using Tensor Flow:

- Input layer corresponding to selected features
- Two hidden layers with ReLU activation
- Output layer using Softmax for binary classification
- Optimizer: Adam
- Loss function: Binary cross-entropy

4.4 Model Training and Evaluation

The dataset was split into training (70%) and testing (30%) sets. Cross-validation and hyper parameter tuning (batch size, learning rate, epochs) were used to optimize model performance. Metrics used include:

- Accuracy
- Precision
- Recall

- F1-Score
- ROC-AUC

IX.Results and Discussion

The proposed ANN model achieved the highest accuracy of approximately **88%**, outperforming traditional models like Logistic Regression, Decision Trees, and SVM. The ANN showed higher AUC-ROC scores, suggesting better discrimination between diabetic and non-diabetic individuals. Deep learning models effectively captured nonlinear interactions among variables, which are often overlooked by simpler algorithms.

Challenges included class imbalance and the small dataset size, which were addressed using resampling techniques and dropout layers to prevent over fitting

X.Conclusion

This study confirms that deep learning models, particularly ANNs, provide a significant improvement in the prediction of diabetes risk among Pima Indian women. The model's high accuracy and adaptability make it a promising tool for early diagnosis and targeted intervention. Future work should focus on integrating more diverse datasets, enhancing model explain ability, and validating results in real-world clinical environments.

.

REFERENCES

[1]Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms, Victor Chang1 • Jozeene Bailey2 • Qianwen Ariel Xu2 • Zhili Sun Neural Computing and Applications (2023) [2] Machine-Learning Based Diabetes Classification and Prediction for Healthcare Applications Victor Chang1 • Jozeene Bailey2 • Qianwen Ariel Xu2 • Zhili Sun3 Hindawi Journal of Healthcare Engineering (2022) [3] Prediction of diabetes disease using machine learning algorithms Monalisa Panda1, Debani Prashad Mishra1, Sopa Mousumi Patro1, Surender Reddy Salkuti2 IAES International Journal of Artificial Intelligence (IJ-AI) (2022). [4] A Novel Architecture for Diabetes Patients' Prediction Using K-Means Clustering and SVMNitin Arora and Sumit Kumar Maitra ,1AnupamSingh 4 ,2Mustafa Zuhaer Nayef Al-Dabagh Hindawi Mathematical Problems in Engineering (2021). [5] Diabetes detection based on learning and deep learning approaches Boon Feng Wee1 ·Saaveethya Sivakumar1·King Hann Lim1·W. K. Wong1· Filbert H. Juwono2 [6] Abdulhadi N, Al-Mousa A (2021) Diabetes detection using machine learning classification methods. In: 2021 International Conference on Information Technology (ICIT). IEEE, p 350–354 [7] Albawi S, MohammedTA, Al-Zawi S (2017) Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). [8] Birjais R, Mourya AK, Chauhan R, Kaur H (2019) Prediction and diagnosis of future diabetes risk: a machine learning approach. [9] Chawla R, Madhu S, Makkar B, Ghosh S, Saboo B, Kalra S et al (2020) Rssdi-esi clinical practice recommendations for the management of type 2 diabetes mellitus 2020. [10] Czmil A, Czmil S, Mazur D (2019) A method to detect type 1 diabetes based on physical activity measurements using a mobile device. [11] HaqAU,LiJP,KhanJ,MemonMH,NazirS,AhmadS, KhanGA,AliA(2020)Intelligentm achinelearning approach for effective recognition of diabetes in e-healthcare using clinical data. [12] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vectormachine," International Journal of Engineering Research and Applications (IJERA) [13] K. M. Orabi, Y. M. Kamal, and T. M. Rabah, "Early predictive system for diabetes mellitus disease," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9728, Springer International Publishing [14] M. Pradhan and G. R. Bamnote, "Design of classifier for detection of diabetes mellitus using genetic programming," in Advances in Intelligent Systems and Computing, vol. 327, Springer International Publishing [15] A. Iyer, J. S, and R. Sumbaly, "Diagnosis of diabetes using classification mining techniques," International Journal of Data Mining & Knowledge Management Process, vol. 5, no. 1, pp. 01–14, Jan. 2015, doi: 10.5121/ijdkp.2015.5101. [16] T. A. Rashid, S. M. Abdullah, and R. M. Abdullah, "An intelligent approach for diabetes classification, prediction and description," in Advances in Intelligent Systems and Computing, vol. 424, Springer International Publishing, 2016. [17] P.

Samant and R. Agarwal, “Machine learning techniques for medical diagnosis of diabetes using iris images,” Computer Methods and Programs in Biomedicine, vol. 157, pp. 121–128, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.004. [18] N. Yilmaz, O. Inan, and M. S. Uzer, “A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases,” Journal of Medical Systems, vol. 38, no. 5, May 2014, Art. no. 48, doi: 10.1007/s10916-014-0048-7. [19] N. Nai-arun and R. Moungmai, “Comparison of classifiers for the risk of diabetes prediction,” Procedia Computer Science, vol. 69, pp. 132–142, 2015, doi: 10.1016/j.procs.2015.10.014. [20] A. Mujumdar and V. Vaidehi, “Diabetes prediction using machine learning algorithms,” Procedia Computer Science, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047 [21] K. Vizhi and A. Dash, “Diabetes prediction using machine learning,” International Journal of Advanced Science and Technology, vol. 29, no. 6, pp. 2842–2852, May 2020, doi: 10.32628/cseit2173107. [22] Barhate, Rahul; Kulkarni, (2018). [IEEE 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Pune, India. (2018.8.16-2018.8.18)] 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) - Analysis of Classifiers for Prediction of Type II Diabetes Mellitus, 1–6. doi:10.1109/ICCUBEA.2018.8697856 [23] Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). doi:10.1109/icaccs51430.2021.94 41935. [24] K. Ateeq and G. Ganapathy, "The novel hybrid Modified Particle Swarm Optimization–Neural Network (MPSO NN) Algorithm for classifying the Diabetes," International Journal of Computational Intelligence Research, vol. 13, no. 4, pp. 595-614, 2017. [25] D. K. Choubey and S. Paul, "GA_RBF NN: a classification system for diabetes," International Journal of Biomedical Engineering and Technology, vol. 23, no. 1, pp. 71- 93, 2017. [26] S. CHALO and İ. B. AYDİLEK, "A New Preprocessing Method for Diabetes and Biomedical Data Classification," Qubahan Academic Journal, vol. 2, no. 4, pp. 6-18, 2022. [27] K. I. Taher, A. M. Abdulazeez, and D. A. Zebari, "Data mining classification algorithms for analyzing soil data," Asian Journal of Research in Computer Science, vol. 8, no. 2, pp. 17-28, 2021. [28] Pustokhina IV, Pustokhin DA, Gupta D, Khanna A, Shankar K, Nguyen GN (2020) An effective training scheme for deep neural network in edge computing enabled Internet of Medical Things (IoMT) systems. IEEE Access 8:107112–107123 [29] Bose, J.S.C., Shankar Kumar, K.R.: Detection of micro classification in mammograms using soft computing techniques. Eur. J. Sci. Res. 86(1), 103–122 (2012). [30] Chaudhuri, A. K., & Das, A. (2020). Variable Selection in Genetic Algorithm Model with Logistic Regression Progression to for Diseases. Prediction 2020 of IEEE International Conference for Innovation in Technology (INOCON). doi:10.1109/inocon50539.2020.9 298372. [31] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, “Disease Prediction by Machine Learning over Big Data from Healthcare Communities,” IEEE Access, vol. 5, no. c, pp. 8869–8879, 2017.[32] K. Lin, M. Chen, J. Deng, M. M. Hassan, and G. Fortino, “Enhanced Fingerprinting and Trajectory Prediction for IoT Localization in Smart Buildings,” IEEE Trans. Autom. Sci. Eng., vol. 13, no. 3, pp. 1294–1307, Jul. 2016 [33] Bose, J.S.C., Shankar Kumar, K.R.: Detection of micro classification in mammograms using soft computing techniques. Eur. J. Sci. Res. 86(1), 103–122 (2012) [34] V. Sangeetha and K Rajesh “Application of data mining Methods and Techniques for diabetes Diagnosis , International Journal of engineering and innovative technology (IJET). [35] Schulz LO, Bennett PH, Ravussin E, Kidd JR, Kidd KK, Esparza J, Valencia ME (2006) Effects of traditional and western envi ronments on prevalence of type 2 diabetes in Pima Indians in Mexico and the US. Diabetes Care 29(8):1866–1871. [36] Smith JW, Everhart JE, Dickson WC, Knowler WC, Johannes RS (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care, pp 261–265 [37] Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T (2019) Current techniques for diabetes prediction: review and case study. Appl Sci 9(21):4604. [38] Pratap Singh R, Javaid M, Haleem A, Vaishya R, Ali S (2020) Internet of Medical Things (IoMT) for orthopaedic in COVID-19 pandemic: roles, challenges, and applications. J Clin Orthop Trauma 11(4):713–717. [39] Cheng D, Ting C, Ho

C, Ho C (2020) Performance evaluation of explainable machine learning on non-communicable diseases. Solid State Technol 63:2780–2793 [40] Iyer A, Jeyalatha S, Sumbaly R (2015) Diagnosis of diabetes using classification mining techniques. Int J Data Min Knowl Managt Process (IJDKP) 5(1):1–14 [41] Mercaldo F, Nardone V, Santone A (2017) Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Procedia Comput Sci 112:2519–252.