MACHINE LEARNING BASED DIABETES CLASSIFICATION AND PREDICTION

D. Suman[1], R.Vishwas[2], B. Sravani[3], M. Bunny[4], B. Rahul[5], B.Veeru[1]

[1]Assistant Professor, Department of CSE, Balaji Institute of Technology & Science, Laknepally, Warangal, India

[2345]BTech Student, Department of CSE, Balaji Institute of Technology and Science, Laknepally, Warangal, India

ABSTRACT

Millions of people worldwide suffer from diabetes, a chronic illness that must be diagnosed early and effectively managed to avoid serious complications. With its sophisticated techniques for precise diabetes prediction and classification, machine learning has become a potent instrument in the medical field.This study evaluates patient health data, such as blood glucose levels, age, BMI, and lifestyle factors, using machine learning algorithms to classify individuals as either diabetic or non-diabetic.The suggested system combines cloud computing for scalable processing with Internet of Things devices for real-time data collection.

1.INTRODUCTION

Diabetes, often known as diabetes mellitus, is a chronic metabolic disease characterized by elevated blood glucose levels. Disease arises when the body is unable to use the insulin it produces or is unable to produce enough of it. The hormone insulin, which is secreted by the pancreas, makes it simpler for cells to absorb glucose for use as fuel. Blood sugar regulation depends on this hormone. Sedentary lifestyles, bad diets, ageing populations, and urbanization have all played a part in the steady increase in the prevalence of diabetes over the past few decades. According to projections from the International Diabetes Federation (IDF), 463 million persons aged 20 to 79 worldwide have diabetes in 2019. If present trends continue, this number is predicted to rise to 700 million by 2045. Diabetes can lead to serious health problems if it is not properly managed. These implications include cardiovascular disease, stroke, kidney disease, nerve damage (neuropathy), visual problems (retinopathy), foot problems, and an increased risk of infection. In addition to being a major cause of death each year, diabetes is the cause of millions of fatalities worldwide. The financial burden Diabetes has a significant financial impact on individuals, families, healthcare institutions, and society as a whole. Direct medical expenditures for care and treatment and indirect costs like missed employment, disability, and early mortality are the two categories of costs

doi: 10.48047/ijiee.2025.15.4.53

associated with diabetes. Global healthcare expenditures associated with diabetes exceeded USD 760 billion in 2019, according to IDF projections. Diabetes prevention's primary objectives are to promote healthy lifestyles that include consistent exercise, a well-balanced diet, weight control, healthcare services, promote healthy lifestyle choices, and raise awareness of illness to the general public. There are several types of diabetes, however the two main ones are: Diabetes type 1: 1 Cause: Unintentional immune system attack and loss of the pancreatic beta cells that produce insulin. Onset: It is typically diagnosed in children and young adults, with symptoms often appearing during early adolescence. Early detection is crucial for managing the condition effectively and preventing long-term complications. Treatment: Insulin pumps or needles must be used to administer insulin continuously, helping to regulate blood sugar levels throughout the day.. Diabetes type 2: Cause: Abnormal blood glucose levels arise when the body produces insufficient amounts of insulin or develops resistance to it. Onset: Adults are more prone to experience it, while children and teenagers can also get it. Treatment: Start with dietary and activity adjustments to enhance lifestyle. Continue taking insulin or oral medicines as necessary[1-23].

2. LITERATURE SURVEY:

Back in 2021, Adeniyi and his buddies came up with this sweet idea. They put together a system where these tiny gadgets you can wear—like a watch or something—team up with fancy tech to keep tabs on your health all the time. It’s like having a little pal who’s always got your back, texting your doctor, “Yo, here’s how they’re doing.” They made it to help people dealing with things like heart trouble or diabetes, and even to stop stuff from getting bad before it’s too late. It’s all about making healthcare less of a hassle and more about you.

Then, in 2023, Rastogi and Bansal had their own cool thing going. They found a way to guess if someone’s headed toward diabetes by digging through a mountain of health details—like playing detective with your medical info. They picked out the big clues, like your age or if your family’s had it, and tweaked their setup to get the predictions just right. It’s like they made a little warning bell for doctors to say, “Heads up, this one might need some extra love soon!”

Both of these are all about how tech can swoop in and make keeping healthy way less of a mystery. How awesome is that teleconsultations [3].Larabi-Marie-Sainte et al. (2019) conducted a comprehensive review of current diabetes prediction techniques, presenting a case study that evaluates the effectiveness of machine learning models. Their work provides valuable insights into feature extraction and algorithmic performance in diabetes prediction

doi: 10.48047/ijiee.2025.15.4.53

[4]. Komi et al. (2017) explored the application of data mining methods in diabetes prediction. Their study discussed various techniques, emphasizing their capability to handle large datasets and improve diagnostic accuracy through advanced analytics [5].Ramanujam et al. (2020) developed a multilingual decision support system for early diabetes detection. Targeting rural Indian populations, their system utilized machine learning techniques to provide accessible and culturally sensitive healthcare solutions [6]. Maan and his team (2020) looked into how machine learning can figure out what might happen with diabetes. They found that different methods can adjust to this challenge, and they couldn't stop talking about how important it is to get the data ready and fine-tune the tools to make everything work better. On the other hand, Samant and Agarwal (2018a) came up with a cool idea: using machine learning to catch diabetes just by looking at someone's iris. It's a new, easy way to check for the disease that might shake things up in the medical world. Later, in 2018b, they took a deeper dive, testing out different ways to classify diabetes with iris images and figuring out what helps some approaches nail it more than others.

Way back in 1996, Quinlan set the stage for decision tree classifiers. He explained how they tick and why they're so great for sorting out medical diagnoses. His work made it clear that these tools bring a straightforward, dependable vibe to healthcare predictions. Saxena (2017) picked up where Quinlan left off, walking us through the nitty-gritty of decision trees—how they come together and make decisions. It's like a friendly guide for anyone wanting to use them to spot diabetes risks.

When it's time to predict who might get diabetes, the models dig into both old and new info. Things like neural networks, team-up strategies, and logistic regression have all gotten a workout here. Everyone keeps saying the same thing: you've got to pick the right details and prep the data properly if you want these models to shine. Tricks like Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) help pinpoint what really drives diabetes risk. And don't skip the basics—filling in gaps in the data and making sure it's all balanced out—because that's what keeps the predictions solid.

Tech-wise, people often talk about cloud computing as the go-to for handling all this info. But lately, edge computing's been stealing the spotlight. It cuts down on the wait time and privacy headaches of the cloud by letting these machine learning tools run right where the data's

doi: 10.48047/ijiee.2025.15.4.53

coming from. That could mean quicker answers and not having to lean so hard on huge, central setup

3.EXISTING SYSTEM

These days, diabetes classification and prediction systems make extensive use of machine learning algorithms as well as traditional statistical methods like logistic regression, decision trees, and support vector machines (SVM). A popular dataset for training these models is the Pima Indians Diabetes Dataset (PIDD), which provides valuable information for developing effective predictive systems.Users can input their health information into web or mobile applications to receive predictions. However, these systems have several challenges, including issues with data quality, where prediction accuracy may be impacted by noisy or missing data. Additionally, the underrepresentation of diabetic cases in the data may lead to biased results due to imbalanced datasets. Furthermore, many of these systems suffer from scalability problems, which limit their ability to process large datasets or manage real-time data streams effectively. Furthermore, the inclusion of of features for real-time monitoring

4. PROBLEM STATEMENT

Traditional diabetes prediction methods have issues with accuracy and scalability. There is an urgent need for a more advanced system that can handle large datasets, make predictions instantly, and integrate with modern technologies like cloud computing and the Internet of Things (IoT) with ease. These traditional approaches may not effectively accommodate the growing volume of health data or deliver the precision required for effective diabetes management. Therefore, developing a more robust solution that leverages contemporary technological advancements is essential for improving diabetes prediction and management.

doi: 10.48047/ijiee.2025.15.4.53

5. PROPOSED SYSTEM

The system being proposed is intended to improve diabetes classification and prediction using sophisticated machine learning algorithms and new technologies. The system will:

- Apply Sophisticated Algorithms: Apply techniques such as Random Forest, XGBoost, and LightGBM to improve the reliability and accuracy of diabetes predictions.

- Make Use of IoT Devices: Employ wearable devices and sensors to obtain real-time health information, allowing ongoing tracking of appropriate health indicators.
- Use Cloud Computing: Implement the system using cloud platforms like AWS or Google Cloud to achieve scalable storage and computing capabilities with efficient data handling and processing.

- Offer User-Friendly Interfaces: Create web and mobile applications that are easy to use, with interfaces through which users can engage with the system and access their health data conveniently.

6. ADVANTAGES

- Improved Accuracy
- Early Detection
- Personalized Predictions
- Automation of Diagnosis
- Real-time Predictions
- Cost Efficiency
- Handling Complex Data
- Better Risk Stratification
- Adaptability
- Detection of Hidden Patterns

doi: 10.48047/ijiee.2025.15.4.53

7. MODULES

The system is composed of several modules that collaborate in enhancing diabetes prediction and classification. These modules are:

-Data Collection Module: It collects health information from different sources, for example, IoT devices, wearable sensors, and electronic health records (EHRs), to capture completeand current data.

-Data Preprocessing Module:Manages missing or incomplete data, normalizes the datasets to make them consistent, and picks the most important features for training the model to maintain the quality and accuracy of the data.

-Model Training Module:Employs machine learning algorithms like Random Forest, XGBoost, and LightGBM to create predictive models for diabetes, ensuring high accuracy and efficiency in predicting the disease.

-Module for Model Evaluation: This module evaluates the quality of various models by comparing important performance metrics like accuracy, precision, recall, and AUC-ROC. It makes sure that only the most accurate and high-performing models are chosen for diabetes prediction.
-User Interface Module:This module offers user-friendly and intuitive web and mobile interfaces, allowing users to input their health information with ease and access real-time predictions, making it a hassle-free and accessible experience.

8. IMPLEMENTATION

- **Data Collection:**
- Assemble a reliable dataset, such as the **Pima Indians Diabetes Dataset**, Kaggle datasets, or hospital records.
- **Data Preprocessing:**

Handle missing values: Eliminate incomplete data points or employ imputation techniques.
Normalize or standardize numerical features: To enhance model training, change the numerical data's scale.

Encode categorical variables: If necessary, transform categorical data into a numerical format.
Divide the dataset: Separate the dataset into training and testing sets, usually in a 70:30 or 80:20 ratio.

- **Feature Selection:**
    - Apply methods like **correlation analysis**, **Principal Component Analysis (PCA)**, or **feature importance** from Random**Model Selection:**

doi: 10.48047/ijiee.2025.15.4.53

- Choose appropriate ML algorithms (e.g., Logistic Regression, Random Forest, SVM, Gradient Boosting, Neural Networks).
- Experiment with multiple models to compare performance.

- **Model Training:**
  - Train the selected models on the training dataset.
  - Use techniques like cross-validation to ensure robustness.

Model Evaluation:
- Use metrics such as F1-score, ROC-AUC, recall, accuracy, and precision to assess models.
Test the model using the test dataset that hasn't been seen yet.

• Hyperparameter tuning:
Use strategies like Grid Search or Random Search to maximize model performance.

•Deployment:
Use frameworks like Flask and FastAPI to deploy the model that performs the best.

• Maintenance and Observation:
Continue to monitor the model's performance in real-world scenarios.
Retrain the model on a regular basis with new data.
Monitor the model's performance in real-world scenarios.

9. METHODOLOGY

The first step in the methodology is data collection, where a reliable and comprehensive dataset is gathered. This dataset typically includes features such as glucose levels, blood pressure, BMI, insulin levels, age, and pregnancy history (for women). Publicly available datasets like the Pima Indians Diabetes Dataset or curated datasets from healthcare institutions are commonly used. Once the data is collected, it undergoes preprocessing to ensure it is clean and ready for analysis. This involves handling missing values through imputation or removal, normalizing or standardizing numerical features to bring them to a common scale, and encoding categorical variables into numerical formats. The dataset is then split into training and testing sets, usually in an 80:20 or 70:30 ratio, to ensure the model can be trained and evaluated effectively.

doi: 10.48047/ijiee.2025.15.4.53

Collecting information—basically putting together a solid and trustworthy set of details—is the first thing you do in this process. This set usually includes things like how old someone is, their blood pressure, body mass index (BMI), insulin and sugar levels, and, for women, info about any pregnancies they've had before. A lot of times, this info comes from carefully organized public collections or health groups, such as the Pima Indians Diabetes Dataset. Getting the collected info ready makes sure it's clean and good to work with. While getting it ready, you turn categories into numbers, deal with any missing pieces by filling them in or tossing them out, and tweak the number-based details so they're all on the same level using something like standardization or normalization. This gets everything lined up for a smooth analysis.

Once the info is ready, the next move is to pick the right tools—think of them as smart recipes—for teaching a computer to learn from it. Some popular ones are simple guesswork models, choice-making trees, big groups of trees working together, line-drawing separators, and power-up methods like XGBoost or LightGBM, plus brain-like networks. Each of these tools gets trained with the practice info, and tricks like splitting the info into chunks for testing—called cross-checking—are used to make sure the tool works well on new, unseen stuff. This cross-checking step keeps the tool from getting too stuck on the practice info by testing how it does on different pieces of it.

After training, the models are evaluated using various performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide insights into how well the model is performing, particularly in terms of its ability to correctly classify individuals with and without diabetes. The testing dataset, which was not used during training, is used to evaluate the model's performance on unseen data. This step ensures that the model is robust and can make accurate predictions in real-world scenarios.

Hyperparameter tuning is then performed to optimize the model's performance. Techniques like Grid Search or Random Search are used to find the best combination of hyperparameters for the selected algorithm. This step is crucial as it fine-tunes the model to achieve the highest possible accuracy and generalization. Once the best-performing model is identified, it is saved for deployment.

doi: 10.48047/ijiee.2025.15.4.53

The final step is deploying the model into a production environment where it can be used to make predictions on new data. This involves creating a user-friendly interface, such as a web or mobile application, where users can input their data and receive predictions. Frameworks like Flask, FastAPI, or Django are commonly used for deployment, and cloud platforms like AWS, Google Cloud, or Azure can be leveraged for scalable and efficient deployment. After deployment, the model's performance is continuously monitored, and it is periodically retrained with new data to ensure it remains accurate and up-to-date. This end-to-end methodology ensures a robust and reliable system for diabetes classification and prediction

10. CONCLUSION

XG Boost, Random Forest, and Light GBM are dependable machine learning techniques for diabetes classification and prediction because they make full use of their advanced capabilities when working with complex datasets. XG Boost offers high accuracy in relation to its gradient boosting framework.. Random Forest excels in handling data distributions with different kinds of distribution and has fewer chances of overfitting. Light GBM excels at processing large- scale data, with relatively low computational requirements. It will further be possible to increase predictive accuracy and reliability of the diagnosis by integrating these models with real-time data collection systems.

REFERENCES

1. Adeniyi, E. A., Ogundokun, R. O., & Awotunde, J. B. (2021). IoMT-Based Wearable Body Sensors Network Healthcare Monitoring System. In G. Marques, A. K. Bhoi, V. H. C. de Albuquerque, & H. KS (Eds.), IoT in Healthcare and Ambient Assisted Living (Vol. 933, pp.
2. 103–121). Springer Singapore. https://doi.org/10.1007/978-981-15-9897-5_6 R. Rastogi, M. Bansal, Diabetes prediction model using data mining techniques (2023)

3. Divya, K., Sirohi, A., Pande, S., & Malik, R. (2021). An IoMT Assisted Heart Disease Diagnostic System Using Machine Learning Techniques. In A. E. Hassanien, A. Khamparia, D. Gupta, K. Shankar, & A. Slowik (Eds.), Cognitive Internet of Medical Things for Smart Healthcare (Vol. 311, pp. 145–161). https://doi.org/10.1007/978-3-030-55833-8_9 Springer International Publishing.

4. 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. Journal of Clinical Orthopaedics https://doi.org/10.1016/j.jcot.2020.05.011 and Trauma, 11(4), 713–717.

doi: 10.48047/ijiee.2025.15.4.53

5. Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R. and Saba, T. (2019). 'Current Techniques for Diabetes Prediction: Review and Case Study', Applied Sciences, vol. 9, no. 21, pp. 4604.
6. Komi, M., Jun Li, Yongxin Zhai, Xianguo Zhang, 2017. Application of data mining methods in diabetes prediction, in: 2017 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, https://doi.org/10.1109/ICIVC.2017.7984706 China, pp. 1006–1010.
7. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" A Brain Tumor Identification and Classification Using Deep Learning based on CNN-LSTM Method" Computers and Electrical Engineering , 101 (2022) 107960
8. Ramdas Vankdothu,.Mohd Abdul Hameed "Adaptive features selection and EDNN based brain image recognition on the internet of medical things", Computers and Electrical Engineering , 103 (2022) 108338.
9. Ramdas Vankdothu,.Mohd Abdul Hameed,Ayesha Ameen,Raheem,Unnisa " Brain image identification and classification on Internet of Medical Things in healthcare system using support value based deep neural network" Computers and Electrical Engineering,102(2022) 108196.
10. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor segmentation of MR images using SVM and fuzzy classifier in machine learning" Measurement: Sensors Journal,Volume 24, 2022, 100440 .
11. Ramdas Vankdothu,.Mohd Abdul Hameed" Brain tumor MRI images identification and classification based on the recurrent convolutional neural network" Measurement: Sensors Journal,Volume 24, 2022, 100412 .
12. Bhukya Madhu, M.Venu Gopala Chari, Ramdas Vankdothu,.Arun Kumar Silivery,Veerender Aerranagula " Intrusion detection models for IOT networks via deep learning approaches " Measurement: Sensors Journal,Volume 25, 2022, 100641
13. Mohd Thousif Ahemad ,Mohd Abdul Hameed, Ramdas Vankdothu" COVID-19 detection and classification for machine learning methods using human genomic data" Measurement: Sensors Journal,Volume 24, 2022, 100537
14. S. Rakesh [a], NagaratnaP. Hegde [b], M. VenuGopalachari [c], D. Jayaram [c], Bhukya Madhu [d], MohdAbdul Hameed [a], Ramdas Vankdothu [e], L.K. Suresh Kumar "Moving object detection using modified GMM based background subtraction" Measurement: Sensors ,Journal,Volume 30, 2023, 100898
15. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Efficient Detection of Brain

doi: 10.48047/ijiee.2025.15.4.53

Tumor Using Unsupervised Modified Deep Belief Network in Big Data" Journal of Adv Research in Dynamical & Control Systems, Vol. 12, 2020.

16. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Internet of Medical Things of Brain Image Recognition Algorithm and High Performance Computing by Convolutional Neural Network" International Journal of Advanced Science and Technology, Vol. 29, No. 6, (2020), pp. 2875 – 2881
17. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima "Convolutional Neural Network-Based Brain Image Recognition Algorithm And High-Performance Computing", Journal Of Critical Reviews,Vol 7, Issue 08, 2020(Scopus Indexed)
18. Ramdas Vankdothu, Dr.Mohd Abdul Hameed "A Security Applicable with Deep Learning Algorithm for Big Data Analysis",Test Engineering & Management Journal,January-February 2020
19. Ramdas Vankdothu, G. Shyama Chandra Prasad " A Study on Privacy Applicable Deep Learning Schemes for Big Data" Complexity International Journal, Volume 23, Issue 2, July-August 2019
20. Ramdas Vankdothu, Dr.Mohd Abdul Hameed, Husnah Fatima " Brain Image Recognition using Internet of Medical Things based Support Value based Adaptive Deep Neural Network" The International journal of analytical and experimental modal analysis, Volume XII, Issue IV, April/2020
21. Ramdas Vankdothu,Dr.Mohd Abdul Hameed, Husnah Fatima" Adaptive Features Selection and EDNN based Brain Image Recognition In Internet Of Medical Things " Journal of Engineering Sciences, Vol 11,Issue 4 , April/ 2020(UGC Care Journal)
22. Ramdas Vankdothu, Dr.Mohd Abdul Hameed " Implementation of a Privacy based Deep Learning Algorithm for Big Data Analytics", Complexity International Journal , Volume 24, Issue 01, Jan 2020
23. Ramdas Vankdothu, G. Shyama Chandra Prasad" A Survey On Big Data Analytics: Challenges, Open Research Issues and Tools" International Journal For Innovative Engineering and Management Research,Vol 08 Issue08, Aug 2019

BIBLIOGRAPHY

I am Racharla Vishwas from the Department of Computer Science and Engineering. Currently, pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on "MACHINE LEARNING BASED DIABETES CLASSIFICATION AND PREDICTION".

doi: 10.48047/ijiee.2025.15.4.53

I am Balasani sravani from the Department of Computer Science and Engineering. Currently, pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on “MACHINE LEARNING BASED DIABETES CLASSIFICATION AND PREDICTION ”.

I am Rahul from the Department of Computer Science and Engineering. Currently, pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on “MACHINE LEARNING BASED DIABETES CLASSIFICATION AND PREDICTION”.

I am More Bunny from the Department of Computer Science and Engineering. Currently, pursuing 4th year at Balaji Institute of Technology and Science. My research is done based on “MACHINE LEARNING BASED DIABETES CLASSIFICATION AND PREDICTION”.

doi: 10.48047/ijiee.2025.15.4.53