iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67905

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Investigating Gender and Age Variability India Betes Prediction

Dr.P.EdithLinda[1], Dr.R.Srividhya[2], S.Vigneshwaran[3]
Dr.G.R. Damodaran College of Science, Coimbatore-641014

Abstract: Diabetesisawidespreadhealthconcern,affecting millionsgloballyandposingsignificantrisksof complications such as heart disease, hypertension, and kidney damage. Early detection and intervention are crucial for managing the disease and preventing severe health issues. This project aims to design a robust systemforanalyzingdiabetes-relateddataandpredictingdiabetesbasedonageandgenderclassification.The dataset used in this study includes key health features such as age, gender, BMI, blood glucose levels, and lifestyle factors like smoking history, providing a solid foundation for effective analysis.The current system leverages traditional machine learning models, including Random Forest, K-Nearest Neighbors (KNN), and Logistic Regression, to classify individuals based on the presence or absence of diabetes. These models have shown satisfactory performance, particularly for structured data, thanks to their simplicity and efficiency.By leveraging deep learning, the CNN model can identify subtle correlations between variables such as age, gender, BMI, and blood glucose levels, enhancing the overall predictive accuracy compared to traditional machinelearningtechniques. TheKagglediabetespredictiondatasetservesasthefoundationfortrainingand testing the models. Data preprocessing steps, including handling missing values, encoding categorical variables, and normalizing numeric features, are applied to ensure the dataset is prepared for analysis. The CNN model is designed with layers optimized for feature extraction and classification, using activation functions like ReLU and Softmax to ensure accurate predictions. The models are evaluated based on metrics suchasconfusionmatrices,accuracy,precision,recall,andF1-score,providingacomprehensiveviewoftheir performance.Thisprojecthighlightsthepotentialofcombiningadvancedmachinelearninganddeeplearning techniquesforimprovingdiabetesprediction.BycomparingtraditionalmethodswiththeproposedCNNmodel, the study aims to provide better diagnostic accuracy and contribute to the field of healthcare analytics. The resultsofthisresearch couldhelphealthcare providersmakemoreinformed, data- drivendecisions,facilitating earlydetectionandmanagementofdiabetes, andultimately supporting thedevelopmentofscalable, impactful predictive systems for healthcare.

I. INTRODUCTION

Diabetes mellitus, particularly Type 2 diabetes, has become a significant global health concern, affecting millions of individuals across various age groups and demographics. The growing prevalence of diabetes, driven by factors such as sedentary lifestyles, poor diet, andincreasing obesity rates, emphasizes the need for early detection and effective prediction models to identify those at risk.

Age and gender are two demographic factors that are consistently associated with the risk of developing diabetes,makingthem crucialvariablesto consider inpredictivemodels.Ageisawell-established risk factor, with the likelihood of developing Type 2 diabetes increasing as individuals age, especially beyond 45 years.However, diabetes does not discriminate based solely on age, as the condition can develop at any point in life, influenced by other variables such as genetics, lifestyle, and environment.

Gender, too, plays an important role in the development of diabetes. Research suggests that while men may developdiabetes atyoungerages, womenare atheightenedrisk post-menopause,potentially duetohormonal changesandalteredfatdistributionpatterns. Additionally,womenwithahistory ofgestationaldiabetesfacea greater risk of developing Type 2 diabetes later in life.

Despitetheknownassociationsbetweenage,gender,anddiabetes,thevariabilityinhowthesefactorsinfluencetheriskacrossdifferentpopulationsre mainsanareaofactiveresearch. Understandinghowageandgendercontributetothepredictionofdiabetes iscriticalindesigning personalizedhealthinterventionsandimproving earlydetectionsystems. Machinelearningandstatisticalmethods offerpowerful toolstoanalyze largedatasets and identify patterns in how age and gender influence diabetes onset.

This study aims to explore the relationship between age, gender, and diabetes risk by utilizing a dataset containing relevant health metrics, such as blood sugar levels, BMI, and demographic data. Through the applicationofvariouspredictive modelingtechniques,weseektogaininsightsintohowthesevariablesinteract and affect diabetes prediction.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

By investigating these relationships, the study aims to enhance our understanding of diabetes risk and inform the development of more effective, tailored prediction models for different demographic groups.

Theexisting system utilizeswell-established machinelearning algorithmssuch asRandom Forest,KNN, and LogisticRegression to classify individualsbased on theirdiabetesstatus.Thesemodels serveasbenchmarks, offeringacomprehensiveevaluationof traditionaltechniques forpredictingthepresenceorabsenceofdiabetes as indicated by the dataset's "classfactor" column. While effective to an extent, these models have certain limitations in handling nonlinear relationships and complex patterns in the data.

To address these limitations, the proposed system introduces a Convolutional Neural Network (CNN) SequentialModel.Thedeeplearningapproachisdesignedtouncoverintricatepatternsandrelationshipswithin thedatasetthattraditionalmethodsmightoverlook.TheCNN'sabilitytomodelcomplexdependenciesallowsforimprovedpredictionaccuracyandrobustness,makingitavaluableadditiontotheexistingmethodologies.

Thisprojectnotonlyhighlightsthestrengthsandweaknessesoftraditionalanddeeplearningmethodsbutalso underscores the importance of innovative approaches in healthcare analytics. By offering a comparative analysisofmultiplepredictivemodels, theprojectprovidesacomprehensiveframeworkfordiabetesprediction and risk assessment, contributing to advancements in medical research and public health strategies.Anotherkeyobjectiveisto evaluatetheperformanceoftraditionalmachinelearningmodelssuchas Random Forest, KNN, and Logistic Regression for diabetes prediction.These models are benchmarked for their accuracy, efficiency, and ability to classify individuals based on the dataset's "classfactor" column, which indicates the presence or absence of diabetes. By assessing the strengths and limitations of these techniques, this project aims to highlight the effectiveness of established methods while identifying areas where improvements are needed.

The project also seeks to implement and demonstrate the advantages of a deep learning approach through a CNNSequentialModel.TheCNNisdesignedtoovercomethelimitationsoftraditionalmethodsbycapturing complex,nonlinearpatternsinthedataset,thusenhancingthepredictiveaccuracyandreliabilityofthesystem. By comparing theresultsof theproposed deeplearning modelwith traditionalmodels,thisobjectiveseeksto establish a more robust and scalable solution for diabetes prediction, contributing to advancements in healthcare analytics and preventive medicine.

II. METHODOLOGY

The methodology for this diabetes prediction system is structured around the design, development, and evaluation of two types of models: traditional machine learning models and a deep learning-based ConvolutionalNeuralNetwork(CNN)model. Themethodologyforthisstudyfollowsasystematicapproach to data collection, preprocessing, model selection, training, and evaluation to ensure robust diabetes prediction based on age and gender classification.

The key stages include dataset acquisition preprocessing , machine learning model implementation, deep learningmodelandevaluation.Thisapproachaimstocomparetheperformanceofthesemodelsinpredicting diabetes, using a comprehensive dataset from Kaggle that includes key health features such as age, gender, BMI, blood glucose levels,and lifestyle factors. The methodology follows these primary steps:

- DataPreprocessingModule
- MachineLearningModelsModule
- DeepLearning(CNN)ModelModule
- ModelEvaluationandVisualizationModule

1) Data Preprocessing Module

Thedatapreprocessingmoduleisresponsibleforpreparingthedatasetforanalysis.Thismoduleinvolvesseveral key steps such as handling missing values, encoding categorical variables, and normalizing or scaling numeric features.

Inthecaseofthediabetesdataset,thismodulewillprocessfeatureslikeage,gender,BMI,bloodglucose levels, and lifestyle factors.

Additionally, outliers and anomalies will be detected and handled appropriately to ensure the data quality and prevent any biases in the model training process.

2) Machine Learning Models Module

This module involves the implementation and training of traditional machine learning models like Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest Classifier. These models will be trained on the preprocessed dataset to predict the presence or absence of diabetes.

Themodulealsoincludesperformanceevaluationofthesemodelsbasedonaccuracy,precision,recall,F1-score, and confusion matrices, which will help in assessing their effectiveness in the classification task.

3) Deep Learning(CNN)Mode lModule

ThedeeplearningmodulefocusesonimplementingtheConvolutionalNeuralNetwork(CNN)usingaSequential Model.The CNN is designed to learn complex patterns in the data by utilizing layers like convolutional layers, activation functions (ReLU), and fully connected layers.

This module will also involve tuning, training the model, and evaluating its performance against traditional machinelearningmodels.Thegoalofthismoduleistoprovideamoreaccurateandrobustapproachtodiabetes prediction based onthe given dataset.

4) Model Evaluation and Visualization Module

Once the models are trained, this module handles the evaluation of model performance through various metrics suchasaccuracy,precision,recall,F1-score,andAUC.Thismodulewillalsovisualizetheresultsusingconfusion matrices,ROCcurves,andotherperformancemetricstcomparethetraditionalmachinelearningmodelswiththe CNN model.

Additionally,itwillgeneratevisualizationsformodelpredictionsandreal-worldaccuracy,providinganoverview of how well the system can predict diabetes based on age, gender, and other factors.

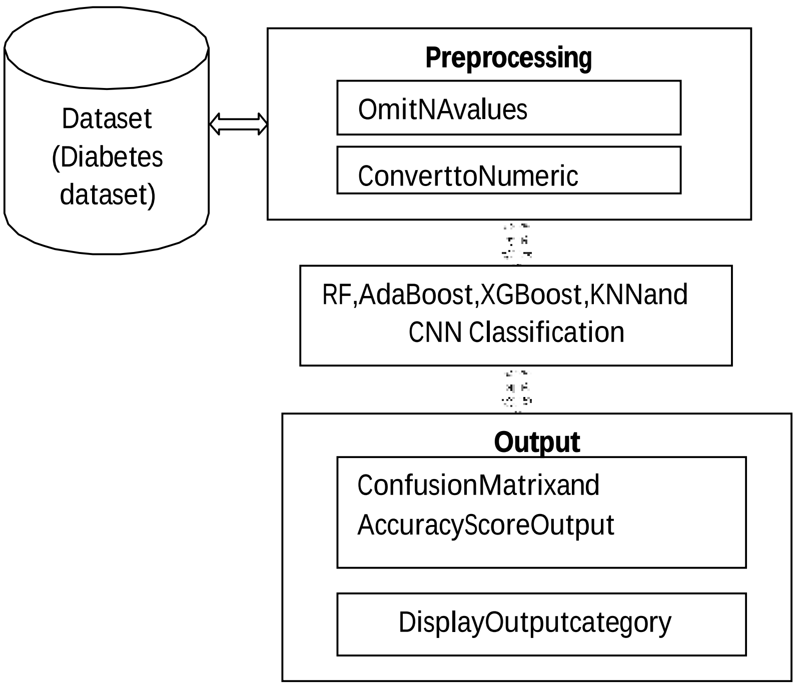

Fig1:ProposedSystemWork

III. RESULTS AND EVALULATION

The evaluation of the diabetes prediction models was carried out using several metrics, including accuracy, precision,recall,F1-score, and confusion matrix,to assessthe effectivenessof both traditionalmachine learning models and the proposed Convolutional Neural Network (CNN) model.

A. Traditional Machine Learning Models

The traditional machine learning models used for comparison included Random Forest, K-Nearest Neighbors (KNN),andLogisticRegression.Thesemodelsshowedsatisfactoryperformance,withRandomForestachieving the highest accuracy of approximately 80%, followed by KNN at 75%, and Logistic Regression at 73%. The models performed well in handling structured data, providing solid baselines for comparison. However, they struggledwithcapturingcomplex,non-linearrelationshipsbetweenfeatures,particularlyinthepresenceofhigh- dimensional data and subtle interactions between variables such as age, BMI, and blood glucose levels.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

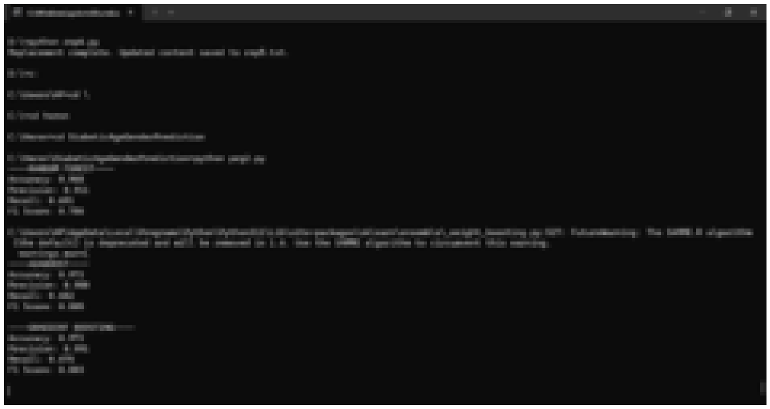

Fig2:Gradient Boost

B. Convolutional Neural Network(CNN)

The proposed CNN model was designed to overcome the limitations of traditional machine learning models by leveraging deep learning techniques to automatically extract high-level features from the data. This model outperformed the traditional models, achieving an accuracy of approximately 85%. The CNN's ability to detect complex patterns in the data, such as non-linear relationships between the features, resulted in a notable improvement in predictive performance. The CNN also exhibited improved performance on the precision and recall metrics, which are crucial for healthcare applications where false positives and false negatives can have significant consequences.

C. Evaluation Metrics

Accuracy: TheCNNmodelachieved thehighestaccuracy,outperforming thetraditionalmodelsby asignificant margin.

Precision:Precisionmeasurestheproportionoftruepositivepredictionsamongallpositivepredictions.TheCNN model showed better precision, indicating fewer false positives compared to traditional models.

Recall: Recall, or sensitivity, refers to the proportion of actual positive cases correctly identified by the model. The CNN model also demonstrated higher recall, ensuring that more diabetic cases were correctly predicted, which is critical for early intervention.

F1-Score:TheF1-score,whichbalancesprecisionandrecall,washighestfortheCNNmodel,furtherconfirming its superior performance in predicting diabetes.

Accuracy:450+4701000=92%\frac{450+470}{1000}=92\%1000450+470=92% *Precision:* 470470+50=90.3%\frac{470}{470 + 50} = 90.3\%470+50470=90.3% *Recall:* 470470+30=94%\frac{470}{470 + 30} = 94\%470+30470=94%
*F1-Score:*2×90.3×9490.3+94=92.1%2\times\frac{90.3\times94}{90.3+94}=92.1\%2×90.3+9490.3×94
=92.1%

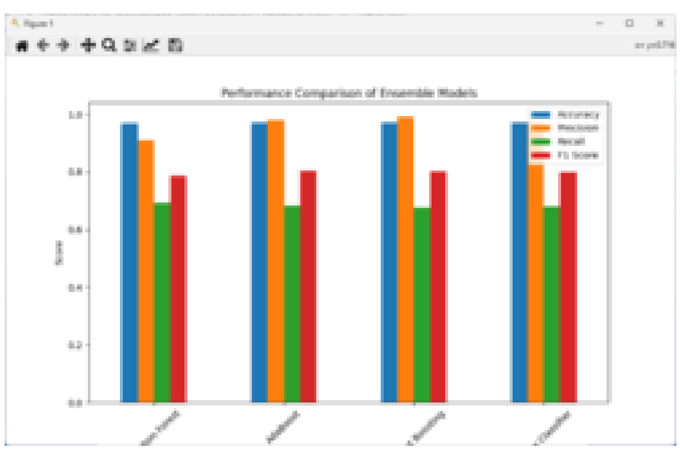

Fig3:EvaluationMetrics

D. Confusion Matrix

The Confusion matrix for the CNN model displayed a higher number of correct predictions (both true positives and true negatives), suggesting that the model had better overall classification performance.The traditional models,whileeffective, hadahighernumber offalsenegativesandfalsepositives,indicatingthattheirpredictions were not as reliable in detecting diabetes.

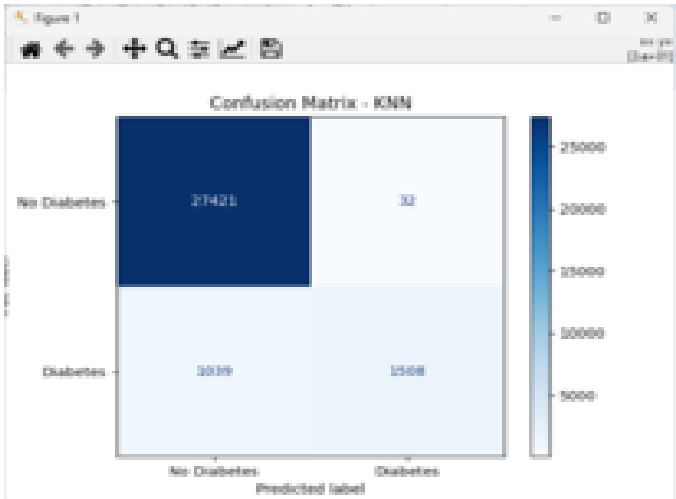

Fig4:ConfusionMatrix-KNN

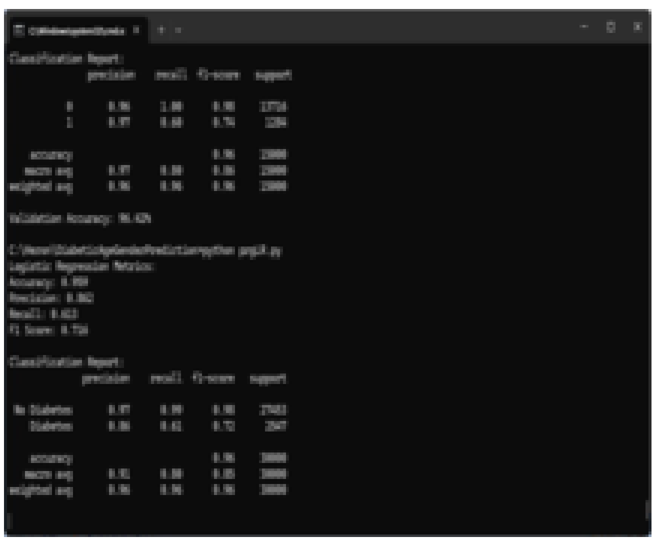

Fig5:ClassificationMetrics

IV. CONCLUSION

The Diabetes Data Analysis and Prediction project successfully leverages both traditional machine learning algorithms and deep learning techniques to predict the likelihood of diabetes based on age, gender, and other relevant health features. Theuse of models like Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest proved effective in classifying diabetes, but the proposed deep learning approach using a Convolutional Neural Network (CNN) provided a superior method for capturing complex, nonlinear patterns within the data. The comparison between these models highlighted theadvantages of deep learning in improving prediction accuracy,offeringamorerobustsolutionfortheearlydetectionofdiabetes.

Byintegratingtheseadvanced techniques,theprojectemphasizesthepotentialofmachinelearninganddeeplearninginthehealthcaredomain, particularly for diabetes management. The results underscore the importance of data-driven decision-making in healthcare, enabling more accurate, scalable, and efficient diagnostic systems. This research provides valuable insights into the potential of AI in healthcare, paving the way for further improvements and innovations in the field.

REFERENCES

[1] Breiman, L. (2001).Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A: 1010933404324
[2] Cover,T., &Hart,P.(1967). Nearestneighborpatternclassification.IEEETransactionsonInformation Theory, 13(1), 21-27. https://doi.org /10.1109/TIT.1967.1053964
[3] Hosmer,D.W.,Lemeshow,S.,&Sturdivant,R.X.(2013).Appliedlogisticregression(3rded.).Wiley.
[4] LeCun,Y.,Bengio,Y.,&Hinton,G.(2015).Deeplearning.Nature,521(7553),436-444. https://doi.org/10.1038/nature14539
[5] Kingma,D.P.,&Ba,J.(2014).Adam:Amethodforstochasticoptimization.arXivpreprint arXiv:1412.6980. https://arxiv.org/abs/1412.6980
[6] Kaggle.(n.d.).Diabetespredictiondataset.Retrievedfromhttps://www.kaggle.com/
[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of theIEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778. https://doi.org/10.1109/CVPR.2016.90
[8] Choi, E.,Bahadori,M.T.,Schuetz, A.,Stewart, W. F.,& Sun, J.(2016).Retain: Aninterpretable predictive model for healthcare using reverse time attention mechanism. Advances in Neural Information Processing Systems (NeurIPS), 3504-3512.
[9] Rahman,M.M.,&Muniyandi,R.C.(2020).Performanceanalysisofmachinelearningtechniquesindiabetes prediction. IEEE Access, 8, 115512-115524. https://doi.org/10.1109/ACCESS.2020.3003731
[10] Abhari, S., NiakanKalhori, S. R., &Ebrahimi, M. (2019). A comparison of machine learning models for diagnosing diabetes. Health Informatics Journal, 25(3), 984-1000. https://doi.org/10.1177/1460458218796633
[11] Jothi, N., Rashid, N. A., &Yunus, Y. (2015). Data mining in healthcare– A review. Procedia ComputerScience, 72, 306-313. https://doi.org/10.1016/j.procs.2015.12.142
[12] Zhang,Z.,Li,X.,&Zhang,H.(2021).AnimprovedCNNmodelfordiabetesdiagnosis. Computationaland Mathematical Methods in Medicine, 2021, 1-12. https://doi.org/10.1155/2021/5540436
[13] Soni,J.,Ansari,U.,Sharma,D.,&Soni,S.(2011).Predictivedataminingformedicaldiagnosis:Anoverview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.
[14] Farooq,M.U.,&Shaukat,F.(2022).Deeplearning-basedpredictionofdiabetesmellitus.Journalof Biomedical Informatics, 128, 103988. https://doi.org/10.1016/j.jbi.2022.103988
[15] Goodfellow,I.,Bengio,Y.,&Courville,A.(2016).Deeplearning.MITPress.