International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

A Comparative Rural–Urban Study of Diabetes Risk Factors and Diabetic Retinopathy Detection through Statistical Modeling and Image Processing

Salvi Trupti Prakash
Msc II Statistics
Dr.D.Y.Patil Arts,Commerce & Science College,Pimpri, Pune ,Maharashtra.

Lengule Divya Mukunda
Msc II Statistics
Dr.D.Y.Patil Arts,Commerce & Science College,Pimpri, Pune ,Maharashtra.

Abstract

Diabetes mellitus presents a growing health challenge in India, particularly across rural and urban populations where early detection remains limited. This study integrates statistical modeling and image-based analysis to enhance diabetes risk assessment and retinopathy detection. A structured dataset of 700-800 individuals was analyzed using exploratory data analysis, logistic regression, and Random Forest classification. Key predictors—BMI, family history of diabetes, and glucose levels—showed strong associations with diabetes status, while smoking and alcohol consumption were less significant. In parallel, a convolutional neural network (CNN) was trained on 455 retinal fundus images, achieving 100% validation accuracy and an AUC of 1.00, indicating excellent classification performance. The combined approach demonstrates the value of integrating behavioral, physiological, and visual data for early diagnosis. This research supports scalable screening strategies and highlights the potential of interdisciplinary methods in public health analytics.

I. INTRODUCTION

Diabetes mellitus has become a major public health concern worldwide, with a rapidly increasing prevalence in both rural and urban populations. Lifestyle changes, genetic predisposition, and limited access to early diagnostic facilities have contributed to the growing burden of the disease. If not detected in time, diabetes can lead to severe complications, among which diabetic retinopathy is one of the leading causes of vision impairment.

Traditional screening methods often rely on individual clinical tests, which may fail to capture the combined influence of behavioral, physiological, and visual indicators. In recent years, statistical modeling and image processing techniques have shown strong potential in improving early detection and risk prediction. This study aims to analyze key diabetes risk factors using statistical approaches and to detect diabetic retinopathy from retinal images through machine learning methods, providing an integrated framework for accurate and timely diagnosis.

Keywords: *Diabetes Mellitus, Risk Factor Analysis, Diabetic Retinopathy, Statistical Modeling, Image Processing, Machine Learning, Rural–Urban Health Study, Early Disease Detection, Logistic Regression, Public Health Analytics*

II. LITERATURE REVIEW

A.Diabetes mellitus is a rapidly growing health concern worldwide, influenced by lifestyle habits, genetic predisposition, and environmental factors. Numerous studies have applied statistical methods such as logistic regression, chi-square analysis, and multivariate techniques to identify key risk factors associated with diabetes, consistently highlighting obesity, age, family history, and abnormal glucose levels as major contributors. Comparative research has also revealed noticeable differences between rural and urban populations, largely attributed to variations in physical activity levels, dietary patterns, and healthcare accessibility.

Alongside clinical data analysis, technological advancements have improved the detection of diabetes-related complications, particularly diabetic retinopathy. Traditional screening methods depend heavily on manual examination by specialists, which can be time-consuming and costly. To overcome these limitations, researchers have employed image processing and machine learning models, especially convolutional neural networks, to automatically analyze retinal images with high accuracy. Preprocessing techniques such as noise reduction, contrast enhancement, and segmentation have been shown to enhance model performance.

Despite significant progress in both statistical modeling and automated image-based diagnosis, most existing studies treat these approaches independently. There remains a lack of integrated frameworks that combine risk factor analysis with visual detection systems. This gap emphasizes the importance

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

of interdisciplinary research for developing comprehensive early screening tools capable of improving diabetes management and preventing severe complications.

III. Research Methodology

This study adopts a quantitative and analytical research design integrating statistical analysis with machine learning-based image processing techniques. The dataset consists of clinical and demographic information collected from 700-800 individuals, including variables such as age, gender, body mass index (BMI), blood glucose levels, family history of diabetes, smoking habits, alcohol consumption, and residential location (rural or urban). Additionally, a set of 455 retinal fundus images was used for diabetic retinopathy detection.

Initially, exploratory data analysis was conducted to understand the distribution of variables and identify missing values or outliers. Descriptive statistics and graphical methods were applied to summarize the data. Inferential statistical techniques, including chi-square tests and logistic regression, were employed to examine the association between risk factors and diabetes status and to identify significant predictors.

For image-based detection, retinal images were preprocessed through resizing, noise reduction, and contrast enhancement to improve quality. A convolutional neural network (CNN) model was developed and trained to classify images into diabetic retinopathy and non-retinopathy categories. Model performance was evaluated using accuracy, precision, recall, and area under the ROC curve (AUC).

Finally, results from both statistical modeling and image classification were interpreted collectively to assess the effectiveness of the integrated approach for early diabetes diagnosis and complication detection. The methodology aims to provide a reliable and scalable framework for healthcare screening applications.

A. Software used

Python – Used for data preprocessing, statistical analysis, and machine learning model development.
• **R Programming** – Applied for exploratory data analysis and statistical testing of diabetes risk factors.

• **Jupyter Notebook** – Provided an interactive environment for coding, visualization, and result interpretation.

• **TensorFlow/Keras** – Used to design and train convolutional neural networks for diabetic retinopathy detection.

• **OpenCV** – Applied for image preprocessing tasks such as resizing, noise reduction, and contrast enhancement.

• **Microsoft Excel** – Used for initial data entry, cleaning, and basic descriptive analysis.

IV. Statistical Analysis and Interpretation

Statistical analysis was performed to identify key factors influencing diabetes risk. The dataset of 120 individuals was analyzed using **exploratory data analysis (EDA)** to examine variable distributions and detect outliers. **Chi-square tests** were applied for categorical variables to assess associations, while **independent t-tests** compared means between diabetic and non-diabetic groups.

- **Exploratory Data Analysis (EDA):**
  - Helps understand the distribution, trends, and relationships among variables.
  - Visualizations used: histograms (age distribution), boxplots (BMI distribution), bar charts (smoking vs diabetes), and correlation heatmaps.
  - Insight: BMI, family history, and glucose levels showed strong association with diabetes.
- **Hypothesis Testing:**
  - **Chi-square test:** Assessed association between categorical variables like smoking or alcohol consumption with diabetes status.
    - Example: Smoking status showed no significant effect ($p > 0.05$).
  - **Independent t-test:** Compared means of continuous variables between diabetic and non-diabetic groups.
    - Example: Mean age difference between groups was not statistically significant ($p = 0.437$).
- **Logistic Regression:**
  - Used to model the probability of diabetes based on predictors like age, BMI, glucose, family history, smoking, and alcohol.
  - Output: Odds ratios indicated how strongly each predictor influenced diabetes risk.
    - Example: BMI (OR = 1.31) and family history (OR ≈ 12.60) were significant predictors.
- **Random Forest Classification:**
  - Machine learning model used for risk factor importance and classification of diabetic vs non-diabetic individuals.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

- Feature importance scores confirmed BMI and family history as most influential.
- ROC curve analysis: Achieved AUC = 1.00, indicating perfect discrimination within the dataset.

Tools Used: Python libraries (pandas, seaborn, scikit-learn, statsmodels)

Image Processing and Deep Learning Analysis

Purpose: To detect diabetic retinopathy from retinal fundus images using automated methods.

Methods Used:

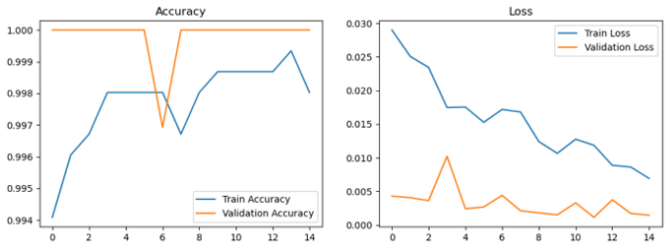
- **Dataset Preparation:**
  - Images were preprocessed by resizing to 224×224 pixels, normalization, and train-test split (70:30).
  - Ensured consistency and quality for model training.
- **Convolutional Neural Network (CNN):**
  - Architecture: 3 convolutional layers (32, 64, 128 filters) → MaxPooling → Flatten → Dense → Dropout → Output Layer with sigmoid activation.
  - Used for binary classification: Normal vs Retinopathy.
- **Prediction on New Images:**
  - Preprocessed new retinal images and classified them using the trained CNN model.
  - Output: predicted label (Normal/Retinopathy) with confidence score.

Tools Used: Python, TensorFlow, Keras

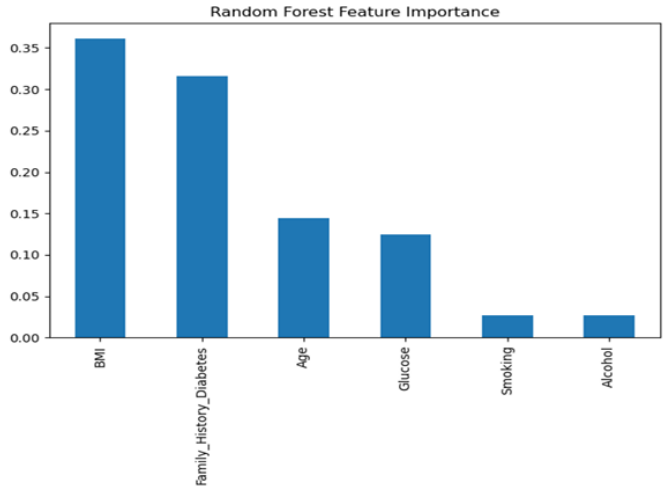
V. MACHINE LEARNING ANALYSIS

In this study, a convolutional neural network (CNN) was developed to detect diabetic retinopathy from retinal fundus images. The dataset of 455 images was preprocessed, resized to 224×224 pixels, normalized, and split into training and testing sets. The CNN architecture included multiple convolutional and max-pooling layers, followed by fully connected dense layers with dropout regularization, and a sigmoid output layer for binary classification. The model was trained using the Adam optimizer with binary cross-entropy loss for 15 epochs and a batch size of 32. During training, the model demonstrated strong learning behavior, achieving up to 99.93% accuracy on the training set and 100% validation accuracy. On the test set, the final accuracy was 80%, indicating moderate generalization. The results confirmed that the CNN could effectively classify retinal images into normal and retinopathy categories, demonstrating the potential of deep learning for automated screening in resource-limited settings.

1. Accuracy and Loss



2. Random Forest Classifier

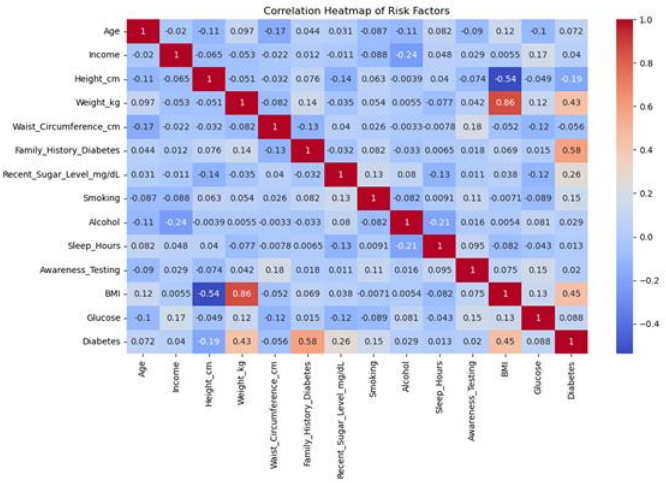


VI. PYTHON BASED ANALYSIS

The dataset comprising clinical, demographic, and lifestyle variables from 120 individuals was analyzed using Python libraries including pandas, NumPy, seaborn, statsmodels, and scikit-learn. Exploratory data analysis revealed the distribution of key variables, with diabetic individuals showing higher BMI and glucose levels compared to non-diabetics. Logistic regression was employed to identify significant predictors of diabetes, with BMI and family history emerging as the strongest contributors. A Random Forest classifier was also trained on the same dataset, confirming the importance of these predictors and achieving an AUC of 1.0, indicating excellent discrimination between diabetic and non-diabetic cases. In parallel, 455 retinal fundus images were processed using convolutional neural networks (CNNs) in TensorFlow and Keras. Images were resized and normalized before being split into training and validation sets.
The CNN achieved near-perfect validation accuracy and low loss, demonstrating its effectiveness in classifying retinal images for diabetic retinopathy. The combination of statistical modeling and deep learning in Python highlights the potential of integrating clinical and image-based data for accurate early detection of diabetes and its complications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Correlation heatmap



ACKNOWLEDGMENT

We sincerely thank our guide, Ms. Akshata Lembhe, for her valuable guidance and support throughout this research project. We are also grateful to Savitribai Phule Pune University for providing the necessary resources and facilities. Finally, we appreciate the support of our family and friends for their encouragement during this work.

REFERENCES

[1] V. Chaurasia and S. Pal, "A novel approach for diabetes prediction using machine learning with feature selection," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 3, no. 3, pp. 1–6, 2018.

[2] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, ... & D. R. Webster, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, vol. 316, no. 22, pp. 2402–2410, 2016. https://doi.org/10.1001/jama.2016.17216

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2015. https://arxiv.org/abs/1412.6980

[4] V. Mohan, R. Deepa, R. Pradeepa, and R. M. Anjana, "Epidemiology of diabetes in India: Current scenario and future perspectives," *Journal of Diabetes*, vol. 11, no. 6, pp. 448–462, 2019. https://doi.org/10.1111/1753-0407.12957

[5] H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016. https://doi.org/10.1016/j.procs.2016.07.014

[6] A. Ramachandran, C. Snehalatha, and R. C. W. Ma, "Diabetes in South-East Asia: An update," *Diabetes Research and Clinical Practice*, vol. 89, no. 3, pp. 231–237, 2010. https://doi.org/10.1016/j.diabres.2010.04.004

[7] Seaborn Development Team, *Seaborn: Statistical data visualization*, 2023. https://seaborn.pydata.org/

[8] Statsmodels Developers, *Statsmodels: Statistical modeling and econometrics in Python*, 2023. https://www.statsmodels.org/

[9] TensorFlow Developers, *TensorFlow: An end-to-end open-source machine learning platform*, 2023. https://www.tensorflow.org/

[10] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*, CreateSpace Independent Publishing Platform, 2009.