Big Data Analytics In Diabetic Management System

A
Project Report

Submitted in partial fulfilment of the
Requirements for the award of the Degree of

BACHELOR OF ENGINEERING

IN

INFORMATION TECHNOLOGY

By

1602-21-737-177 C. Shreya Sree
1602-21-737-315 A. Hasitha

Under the guidance of

Dr. M. Neelakantappa

Associate Professor

Department of Information Technology
Vasavi College of Engineering (Autonomous)
ACCREDITED BY NAAC WITH 'A++' GRADE
(Affiliated to Osmania University)
Ibrahimbagh, Hyderabad-31
2025

Vasavi College of Engineering (Autonomous)

ACCREDITED BY NAAC WITH 'A++' GRADE

(Affiliated to Osmania University)

Hyderabad-500 031

Department of Information Technology

DECLARATION BY THE CANDIDATE

We, **Chintapenta Shreya Sree** and **Amaravadi Hasitha** bearing hall ticket numbers, **1602-21-737-177** and **1602-21-737-315** hereby declare that the project report entitled **Big Data Analytics In Diabetic Management System** under the guidance of **Dr. M. Neelakantappa , Associate Professor**, Department of Information Technology, Vasavi College of Engineering, Hyderabad, is submitted in partial fulfilment of the requirement for the award of the degree of **Bachelor of Engineering** in **Information Technology**

This is a record of bonafide work carried out by us and the results embodied in this project report have not been submitted to any other university or institute for the award of any other degree or diploma.

Chintapenta Shreya Sree
1602-21-737-177

Amaravadi Hasitha
1602-21-737-315

Vasavi College of Engineering (Autonomous)

ACCREDITED BY NAAC WITH 'A++' GRADE

(Affiliated to Osmania University)

Hyderabad-500 031

Department of Information Technology

BONAFIDE CERTIFICATE

This is to certify that the project entitled **Big Data Analytics In Diabetic Management System** being submitted by **C. Shreya Sree** and **A.Hasitha** bearing **1602-21-737-177** and **1602-21-737-315** in partial fulfilment of the requirements for the award of the degree of Bachelor of Engineering in Information Technology is a record of bonafide work carried out by them under my guidance.

Dr. M. Neelakantappa
Associate Professor
Internal Guide

Dr. K. Ram Mohan Rao
Professor & HOD, IT

External Examiner

ACKNOWLEDGEMENT

The satisfaction that accompanies that the successful completion of the Main project would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

It is with immense pleasure that we would like to take the opportunity to express our humble gratitude to **Dr. M. Neelakantappa , Associate Professor, Information Technology** under whom we executed this project. We are also grateful to **Dr. S.K. Chaya Devi, Associate Professor, Information Technology** for his/ her guidance. Their constant guidance and willingness to share their vast knowledge made us understand this project and its manifestations in great depths and helped us to complete the assigned tasks.

We are very much thankful to **Dr. K. Ram Mohan Rao, Professor & HOD, Information Technology,,** for his kind support and for providing necessary facilities to carry out the work.

We wish to convey our special thanks to **Dr. S.V.Ramana, Principal** of **Vasavi College of Engineering and Management** for providing facilities. Not to forget, we thank all other faculty and non-teaching staff, who had directly or indirectly helped and supported me in completing my project in time.

Abstract

This project presents a detailed study focused on predicting diabetes status using a combination of machine learning (ML) and deep learning (DL) models, applied to the BRFSS 2021 diabetes dataset. The dataset includes diverse health-related indicators such as BMI, physical activity, general health, and lifestyle factors, with the target variable Diabetes_012 categorized into three classes: 0 (No Diabetes), 1 (Pre-diabetes), and 2 (Diabetes). The main goal of the study was to build accurate classification models capable of predicting an individual’s diabetes status based on these features.The data preprocessing stage involved handling missing values, separating the target variable from the input features, and splitting the dataset into training and testing sets in an 80:20 ratio. For multi-class classification, encoding was performed, and the One-vs-Rest (OvR) strategy was used to evaluate model performance through AUC (Area Under the Curve) scores. Several traditional ML models were implemented, including Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, and Naive Bayes. These models were assessed using accuracy, confusion matrices, classification reports, and ROC curves.In addition to the ML models, an Artificial Neural Network (ANN) was developed using TensorFlow's Keras API. The ANN comprised two hidden layers with 64 and 32 neurons activated using the ReLU function, followed by a softmax output layer for multi-class classification. The model was optimized using the Adam optimizer and trained with early stopping to prevent overfitting.Among all models, ensemble techniques like Random Forest and Gradient Boosting performed exceptionally well, thanks to their ability to capture complex feature interactions and avoid overfitting. The ANN also demonstrated strong predictive capabilities, especially for non-linear patterns. In conclusion, this study proves the effectiveness of ML and DL in diabetes prediction, offering potential for real-world applications in preventive healthcare and patient monitoring systems.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviations	**Full Forms**
BDA	Big Data Analytics
ML	Machine Learning
MLLib	Machine Learning Library
HER’s	Electronic Health Records
BRFS	Behavioral Risk Factors Surveillance System
MSE	Mean Squared Error
BD	Big Data
IoHT	Internet of Health Things
MCPS	Medical Cyber-Physical Systems
LSCM	Logistics and Supply Chain Management
BDBA	Big Data Business Analytics
SCA	Supply Chain Analytics
DQN	Deep Q-Network
HIE	Health Information Exchanges
T1DM	Type 1 Diabetes Mellitus
T2DM	Type 2 Diabetes Mellitus
AI	Artificial Intelligence
MA	Management accounting
ANN	Artificial Neural Networks
EDA	Exploratory Data Analytics

1. INTRODUCTION

Diabetes has emerged as one of the most pressing public health challenges of the 21st century, affecting millions globally and significantly burdening healthcare systems. Characterized by elevated blood glucose levels due to insulin deficiency or resistance, diabetes has seen a sharp rise in prevalence. According to the World Health Organization, the global diabetic population increased from 108 million in 1980 to over 537 million in 2021 and is projected to exceed 783 million by 2045. In India, which is often referred to as the "Diabetes Capital of the World," the number of diagnosed cases is expected to rise from 77 million in 2021 to over 134 million by 2045, driven by factors such as sedentary lifestyles, unhealthy diets, and limited awareness of preventive healthcare.

This project aims to leverage Big Data Analytics (BDA) and Machine Learning (ML) techniques to enhance early detection and management of diabetes. Using the BRFSS 2021 dataset, which contains extensive health indicators like BMI, age, physical activity, and lifestyle habits, several predictive models were trained and tested. These include Decision Trees, Random Forest, Gradient Boosting, Logistic Regression, Naive Bayes, and an Artificial Neural Network (ANN). Performance was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC curves.

Ensemble models like Random Forest and Gradient Boosting showed high accuracy, while the ANN performed well in capturing complex patterns. The project highlights the importance of integrating BDA into healthcare systems, enabling real-time risk prediction, personalized interventions, and better resource planning. By analyzing large-scale structured health data, such systems support proactive disease management and reduce long-term healthcare costs. Ultimately, the application of data science and ML in diabetes care can revolutionize public health strategies, emphasizing early diagnosis, prevention, and individualized treatment.

1.1. Problem Statement – Overview

Diabetes is a growing global health concern, requiring early detection and effective management to prevent severe complications. Traditional healthcare systems struggle to analyze the vast and complex data generated daily, leading to delayed diagnosis and inefficient treatment strategies. Existing methods lack scalability, real-time processing, and predictive accuracy, limiting their ability to support personalized patient care. This study addresses these challenges by leveraging Big Data Analytics (BDA) and Machine Learning (ML) to enhance diabetes prediction and healthcare decision-making. By integrating scalable data processing frameworks, the proposed approach ensures efficient disease detection, optimized patient management, and improved clinical outcomes while addressing privacy and ethical concerns.

1.2. Motivation

Diabetes is a growing global health concern, affecting millions of people and straining healthcare systems worldwide. Traditional methods of managing and predicting diabetes often lack precision due to the vast and complex nature of medical data. The rise of Big Data Analytics (BDA) presents an opportunity to transform diabetes management by leveraging advanced computing techniques to analyze vast datasets efficiently. This study aims to harness machine learning and data-driven methodologies to enhance healthcare decision-making, ensuring timely diagnosis, personalized treatment, and improved patient outcomes. By integrating big data with healthcare analytics, this research provides a framework for utilizing machine learning models such as Logistic Regression and Random Forest to optimize diabetes prediction. The findings highlight the potential of BDA in reducing healthcare costs, supporting telemedicine, and guiding policy decisions. This study is a step towards a data-driven healthcare system that improves patient care and disease management strategies.

1.3. Scope & Objectives of the Proposed Work

Scope of the Proposed Work

The study explores the integration of **Big Data Analytics (BDA)** and **Machine Learning (ML)** techniques to enhance **diabetes management and healthcare decision-making**. The scope of this research includes:

- **Utilizing BDA in Healthcare:** Implementing data-driven approaches to improve diagnosis, treatment, and patient monitoring.

- **Improving Decision-Making:** Supporting healthcare professionals with data-driven insights for personalized treatment.
- **Integrating Machine Learning Models:** Comparing multiple ML models (Logistic Regression, Decision Trees, Random Forest, etc.) for **diabetes prediction** and selecting the most effective approach.
- **Addressing Healthcare Challenges:** Tackling issues related to early diagnosis, treatment optimization, and healthcare cost reduction.
- **Ensuring Data Privacy and Ethical Considerations:** Implementing robust governance measures to handle patient data responsibly.

Objectives of the Proposed Work

1. **Develop a Big Data Analytics Framework:** Establish a systematic approach for integrating **ML algorithms** into **diabetes healthcare systems**.
2. **Improve Early Detection of Diabetes:** Utilize **ML techniques** to analyze large datasets and predict diabetes risk factors more accurately.
3. **Optimize Machine Learning Model Performance:** Evaluate and compare models (Logistic Regression, Decision Tree, Random Forest, etc.) to determine the most effective algorithm for diabetes classification.
4. **Enhance Healthcare Decision-Making:** Provide **real-time insights** and predictive analysis to assist healthcare professionals in making informed decisions.
5. **Reduce Healthcare Costs and Improve Efficiency:** Utilize **predictive analytics** to minimize unnecessary treatments, hospital visits, and medical expenses.
6. **Facilitate Telemedicine and Remote Monitoring:** Enable real-time patient monitoring and virtual healthcare solutions through big data-driven insights.
7. **Address Ethical and Privacy Concerns:** Implement data security measures to protect sensitive patient information while leveraging big data techniques.

This research aims to revolutionize **diabetes management** by integrating **Big Data Analytics and ML**, ultimately leading to **better patient care, cost reduction, and enhanced healthcare decision-making**.

1.4. Organization of the Report

This report is structured to provide a clear and comprehensive exploration of Big Data Analytics (BDA) in the context of diabetes prediction and healthcare decision-making. Each chapter builds progressively on the previous, ensuring logical development from foundational knowledge to experimental validation. The organization is as follows:

Chapter 1: Introduction

Describes the healthcare burden posed by diabetes, with a focus on global and Indian contexts.

Introduces Big Data Analytics and Machine Learning (ML) as tools to address these challenges.

Defines key concepts and highlights the potential of integrating Electronic Health Records (EHRs) and health indicators.

Chapter 2: Literature Survey

Reviews past research involving BDA and ML in healthcare.

Compares different approaches, datasets, and models used in diabetes prediction.

Identifies limitations in current methodologies and areas for improvement.

Chapter 3: Proposed System

Details the methodology used, including both classical ML models and a deep learning (ANN) approach.

Describes the architecture of the system and functional modules.

Provides a full description of the BRFSS 2021 dataset and its features.

Chapter 4: Experimental Setup and Results

Outlines the technical specifications, tools, and software used .

Discusses preprocessing steps, model training, and evaluation metrics.

Presents confusion matrices, ROC curves, and performance comparison of models.

Chapter 5: Conclusion and Future Scope

Summarizes findings, including the best-performing models.

Suggests future enhancements like deep learning integration and multi-dataset testing.

Discusses scalability, real-time applications, and telemedicine potential.

References and Appendix

Lists all academic sources and datasets cited.

Includes code snippets and additional figures used in the analysis.

2. LITERATURE SURVEY

- **Karatas et al. [1]** provide a comprehensive review of the applications of Big Data (BD) in the context of Industry 4.0 within healthcare systems. Their study examines how technologies such as the Internet of Health Things (IoHT), Medical Cyber–Physical Systems (MCPS), machine learning (ML), and cloud computing are revolutionizing healthcare services. The authors categorize literature based on application domains like disease-specific analytics, patient data privacy, wearable technologies, and healthcare quality measurement. They emphasize the role of BD in predictive diagnosis, patient monitoring, and personalized treatment. The paper also explores challenges like data security and integration, concluding with a future outlook advocating for robust infrastructure and policy support to optimize BD applications in smart healthcare systems.

- **Palanisamy et al. [2]** explore the transformative potential of big data analytics in the development of healthcare frameworks, focusing on how diverse data sources—ranging from electronic health records to social and behavioral data—can be integrated to improve healthcare services. Their study categorizes healthcare stakeholders (patients, practitioners, hospital operators, insurers, researchers) and aligns them with corresponding data types and analytical methods. The paper reviews numerous big data frameworks and architectural models designed to process structured, semi-structured, and unstructured data to deliver personalized, scalable, and predictive healthcare solutions. A significant emphasis is placed on the role of machine learning, real-time analytics, and data integration tools in building robust and secure healthcare ecosystems. The paper concludes by detailing a range of big data tools suited for tasks like stream processing, visual analytics, and secure data sharing, thereby providing a roadmap for designing effective patient-centric healthcare architectures.

- **Nazir et al. [3]** conducted a systematic and in-depth review of big data management and analytics in healthcare, focusing on the challenges and scientific programming solutions involved. The paper organizes 127 primary studies from 2015 to early 2020 to identify significant healthcare data features, such as blood sugar, pulse rate, and respiratory rate, that are used for disease detection and patient monitoring. It outlines the evolution of data-driven healthcare systems through the adoption of IoT devices, wearable technology, and electronic health records, emphasizing the exponential growth in medical data and its complexity. Key contributions include the identification of state-of-the-art big data techniques for storage, real-time processing, and analytics—such as classification, clustering, and regression—for predictive healthcare and personalized treatment. The paper also highlights applications of big data in clinical decision-making, precision medicine, and public health, providing a structured methodology for evaluating research quality and trends. This work serves as a foundational resource for designing intelligent, data-centric healthcare infrastructures.

- **Wang et al. [4]** provide an in-depth examination of big data business analytics (BDBA) applications in logistics and supply chain management (LSCM), framing the concept as Supply Chain Analytics (SCA). Their work systematically classifies existing literature by the type of analytics (descriptive, predictive, prescriptive) and the operational focus (strategic or tactical). The authors propose a five-level SCA maturity framework—functional, process-based, collaborative, agile, and sustainable—to assess the evolution and capabilities of SCA within organizations. They emphasize the strategic integration of BDBA to enhance decision-making in areas like demand forecasting, procurement, production scheduling, inventory control, and logistics optimization. Additionally, they explore advanced analytics techniques (e.g., regression, simulation, optimization) and how these enable real-time, evidence-based decisions across supply networks. Their study calls for a holistic adoption of SCA as a strategic asset for creating integrated enterprise analytics, stressing the role of organizational culture and leadership in driving analytics maturity and sustainability.

- **Wang et al. [5]** present a structured framework for understanding the capabilities and benefits of big data analytics (BDA) in healthcare organizations. Drawing from 26 healthcare implementation cases, they identify five core BDA capabilities: analytical capability for patterns of care, unstructured data analytical capability, decision support capability, predictive capability, and traceability. These capabilities span the entire data lifecycle—from capture and aggregation to analysis and decision-making. The authors propose a layered big data architecture tailored for healthcare, emphasizing data governance, real-time analytics, and interoperability across diverse data sources. The study further classifies the benefits of BDA into five dimensions: IT infrastructure, operational, organizational, managerial, and strategic. These include cost reductions, improved clinical decision-making, enhanced patient monitoring, and support for strategic growth initiatives. Their research addresses a critical gap by highlighting not just technological but also organizational enablers of BDA, offering actionable strategies such as fostering data-sharing cultures and integrating cloud-based solutions.

- **Bebortta et al. [6]** propose the DeepMist framework, a novel architecture integrating deep learning with mist computing to manage large-scale healthcare data, particularly for heart disease prediction. The framework utilizes the Deep Q-Network (DQN) algorithm to process data locally on mist computing nodes, reducing computational load on IoT devices while ensuring low latency and energy efficiency. DeepMist functions through a layered architecture consisting of IoT, mist, fog, and cloud layers, and demonstrates intelligent decision-making by offloading and processing tasks dynamically across these layers.The study reports a heart disease prediction accuracy of 97.67%, outperforming existing models like Q-Reinforcement Learning (QRL) and Deep Reinforcement Learning (DRL) in terms of precision, recall, F-measure, energy consumption, and processing delay. Energy use is minimized to 32.10 mJ, and average delay reduced to 2.80 ms, highlighting the model's real-time efficiency. The DeepMist architecture is validated through simulation on the Cleveland heart disease dataset using the iFogSim toolkit. The authors argue that mist computing, positioned closer to data sources, presents a promising paradigm for real-time, scalable, and intelligent healthcare analytics.

- **Hasan et al. [7]** introduce a robust machine learning framework for early diabetes prediction using ensemble techniques on the Pima Indian Diabetes (PID) dataset. The framework tackles critical data challenges such as missing values, outliers, and imbalanced class distribution through a meticulous preprocessing pipeline that includes outlier rejection, mean-based imputation, standardization, feature selection, and K-fold cross-validation.The study evaluates multiple classifiers—k-NN, Decision Tree, Random Forest, Naive Bayes, AdaBoost, XGBoost, and Multilayer Perceptron (MLP)—and then proposes a weighted ensemble classifier where model contributions are scaled by their AUC scores. This ensembling, particularly using AdaBoost and XGBoost, yielded the highest performance, with an AUC of 0.950, specificity of 0.934, and diagnostic odds ratio of 66.234. The model significantly outperformed prior approaches, indicating its effectiveness in delivering accurate, real-time diabetes predictions.Their extensive comparative analysis under consistent experimental settings validates the superiority of the proposed approach, especially in handling complex medical datasets with limited labeled examples and high feature variability.

- **Ahmed et al. [8]** present a comprehensive systematic literature review (SLR) of Big Data Analytics (BDA) in healthcare, covering a ten-year span from 2013 to 2023. The study synthesizes findings from 180 scientific publications and examines healthcare BDA from multiple dimensions: ecosystem design, applications, frameworks, data sources, challenges, and implications. The paper explores how BDA improves diagnostic accuracy, facilitates personalized treatments, enhances patient outcomes, and reduces healthcare costs. A novel taxonomy is introduced to classify data fusion and multimodal data handling techniques, particularly for complex healthcare environments involving diverse data types like clinical notes, sensor readings, medical imaging, and genomics.The authors emphasize the significance of multimodal analytics, predictive and prescriptive modeling, digital health infrastructures, and privacy-preserving mechanisms such as de-identification and blockchain. They also provide insights into the Healthcare Informatics and Analytics (HCI&A) life-cycle and highlight how emerging technologies (e.g., NLP, AI, IoT) are reshaping the healthcare landscape. The paper outlines open research challenges—interoperability, ethical considerations, resource constraints—and proposes directions for future study. By integrating data governance, secure architecture, and machine learning frameworks, the authors envision a holistic, real-time, patient-centric healthcare ecosystem driven by big data.

- **Hussain et al. [9]** Hussain et al. (2023) offer a detailed systematic review focusing on the integration of Big Data Analytics (BDA) in clinical decision-making within healthcare systems. The study reviews 185 research articles (2012–2023) and categorizes insights into the application of BDA for improving electronic health records (EHRs), medical diagnostics, personalized treatments, and predictive modeling. A central theme is the use of big data tools and technologies (e.g., Apache Hadoop, Spark, Hive, MapReduce) and their roles in storing, processing, querying, and analyzing massive healthcare datasets.The authors explore key BDA challenges in healthcare, including data heterogeneity, integration, security, and the handling of unstructured data (e.g., clinical notes, medical imaging). They also introduce a 10Vs framework—Volume, Velocity, Variety, Veracity, Value, Variability, Visualization, Validity, Vulnerability, and Volatility—as the lens for understanding the complexities of healthcare big data. The review highlights real-world applications of BDA such as early disease detection, fraud detection, population health management, and remote patient monitoring.Significantly, the paper stresses the need for robust infrastructure and skilled personnel, alongside ethical data governance, to fully harness BDA's transformative potential in patient-centric care and operational efficiency. The work concludes with strategic insights into how healthcare providers can leverage BDA for optimized outcomes and cost reduction through scalable, cloud-based analytics platforms.

- **Aceto et al. [10]** offer a comprehensive review in their paper "The Role of Information and Communication Technologies in Healthcare: Taxonomies, Perspectives, and Challenges", identifying the transformative impact of ICTs in the healthcare sector. Their research presents an extensive classification of ICT-based healthcare paradigms, including e-health, mobile health, personalized health, and smart health. By analyzing over 300 scholarly works, the authors highlight the enabling technologies—such as IoT, cloud computing, big data analytics, and artificial intelligence—that underpin these paradigms. They emphasize how these technologies are reshaping healthcare delivery through enhanced data management, improved remote monitoring, and personalized patient services, while also addressing ongoing challenges related to security, privacy, and interoperability (Aceto et al. 125–154).

- **Dash et al. [11]** explore the transformative potential of big data in healthcare, emphasizing its role in modernizing patient care, enabling personalized medicine, and improving clinical outcomes. In their review, they discuss the core sources of healthcare big data, including electronic health records (EHRs), biomedical imaging, IoT devices, and "omics" technologies such as genomics and transcriptomics. The paper highlights how advanced analytics, artificial intelligence (AI), and machine learning (ML) are central to extracting actionable insights from massive, heterogeneous datasets. The authors also detail the architecture and utility of key computational frameworks like Hadoop and Apache Spark in managing and processing healthcare data. Additionally, they address ongoing challenges, such as data heterogeneity, lack of standardization, and security risks. With the integration of AI-based decision systems and quantum computing, the paper envisions a future where healthcare becomes more predictive, cost-efficient, and tailored to individual patient profiles (Dash et al. 1–25).

- **Wang et al. [12]** presents a forward-looking perspective on biomedical big data analytics, focusing on enabling patient-centric and outcome-driven precision health. In her keynote address, she emphasizes the challenges and opportunities involved in managing multi-modal and multi-scale biomedical data generated by next-generation sequencing, -omics technologies, mobile health sensors, and imaging platforms. Wang outlines the biomedical informatics pipeline, which includes data quality control, feature extraction, knowledge modeling, decision-making, and action-taking through feedback. The paper also explores methodological advances such as data integration, case-based reasoning for individualized care, and real-time analytics through streaming data for mobile health applications—including conditions like Sickle Cell Disease, asthma, and diabetes. Furthermore, Wang addresses the workforce gap in biomedical data science and highlights initiatives like MOOCs and community engagement as means to educate stakeholders and scale up expertise. Her work underlines the essential role of informatics in achieving predictive, preventive, personalized, patient-centric, and precise (5P) healthcare (Wang 1–2).

- **Raghupathi et al. [13]** provide a foundational exploration of the promise and potential of big data analytics in healthcare. They outline how digitization of vast health data—from electronic health records and clinical decision support systems to social media and biometric sensors—presents new opportunities for enhancing healthcare outcomes while reducing costs. Their paper introduces a conceptual framework for big data analytics in healthcare, emphasizing the "4 Vs": volume, velocity, variety, and veracity. The authors discuss how advanced analytical tools, such as Hadoop and NoSQL platforms, enable large-scale data processing to support evidence-based decision-making, predictive modeling, and real-time monitoring. Applications include disease surveillance, fraud detection, population health management, and personalized treatment protocols. They cite real-world implementations, such as Kaiser Permanente and Columbia University Medical Center, which demonstrate how big data has led to earlier diagnoses, improved patient care, and cost reductions. Despite the promise, the paper also acknowledges major challenges such as data quality, standardization, privacy, and the steep learning curve associated with new analytics platforms (Raghupathi and Raghupathi 1–10).

- **Kankanhalli et al. [14]** introduce a special section on big data and analytics in healthcare by highlighting the rising relevance of data-driven healthcare in response to aging populations, lifestyle changes, and the digitization of health records. The paper explains the defining characteristics of big data—volume, variety, and velocity—and details the various sources of healthcare data, including EHRs, medical sensors, social media, genomics, and clinical research. The authors emphasize the shift toward evidence-based medicine and the potential of predictive analytics in transforming healthcare by enabling early diagnosis, optimizing treatment plans, and improving patient outcomes. Furthermore, they address the multifaceted challenges that accompany the implementation of big data analytics, ranging from technological integration and data heterogeneity to organizational resistance, privacy concerns, economic considerations, and evolving policy frameworks. The paper sets the stage for further research by emphasizing the need for interdisciplinary approaches and highlights two studies featured in the special section: one leveraging support vector machines for automating medical systematic reviews, and another exploring psychological factors influencing healthcare information protection behaviors among HIS users

- **George et al. [15]** provide a comprehensive survey on the application of big data analytics to predict diabetes using supervised classification methods. Recognizing diabetes mellitus as one of the most prevalent non-communicable diseases globally, the paper explores how large-scale data analysis can uncover patterns and correlations among variables such as age, blood pressure, BMI, insulin levels, and glucose concentration. The authors compare the effectiveness of multiple supervised learning algorithms—such as Support Vector Machines (SVM), Random Forest, Naïve Bayes, and K-Nearest Neighbors (K-NN)—in terms of predictive accuracy. Among these, SVM and Random Forest showed superior performance. The study also outlines the process of predictive analytics in healthcare, emphasizing the role of preprocessing, algorithm selection, and the interpretation of results. The dataset used for analysis includes features like plasma glucose, skin thickness, insulin levels, and number of pregnancies. Additionally, the paper discusses how social media data can contribute to diabetes awareness and prediction, noting varying levels of awareness across age groups and professions. By integrating big data analytics into predictive healthcare, the authors advocate for proactive diabetes management and cost-effective treatment strategies

- **Eswari et al. [16]** propose a predictive methodology to analyze diabetic data using big data technologies, particularly within a Hadoop/MapReduce environment. Recognizing diabetes mellitus (DM) as a significant non-communicable disease in developing nations like India, the authors emphasize the need for structuring vast unstructured health data to improve accessibility, affordability, and patient care. The proposed system integrates data from diverse sources such as EHRs, prescriptions, diagnostic reports, and social media, using

Hadoop’s parallel processing capabilities for effective analysis. The model identifies diabetes types, associated complications, and appropriate treatments by employing predictive analytics that includes classification, clustering, and pattern discovery techniques. The authors also highlight the importance of Health Information Exchanges (HIE) for seamless, secure access to integrated patient data. Ultimately, their system demonstrates how big data infrastructure can enable early detection, facilitate personalized treatment strategies, and reduce healthcare costs, especially for underserved rural populations

- **Kolesnichenko et al. [17]** present a novel approach using big data analytics and advanced data visualization to study inpatient flow among patients with Type 1 Diabetes Mellitus (T1DM). Utilizing a cloud-based medical information system (qMS), they processed large-scale clinical data from 862 patients across multiple Russian hospitals using a range of analytic techniques—cluster analysis, graph theory, Boolean algebra with Gray code, and 3D visualization. Their study constructs a "pathogenetic continuum" of T1DM, identifying five patient models and uncovering correlations between diabetes progression and complications such as lung cancer, cardiovascular issues, and osteoporosis. A central finding is the potential role of parathyroid hormone-related protein (PTHrP) in the disease's multi-organ pathogenesis and its prospective use in pharmacological treatment. Through multidimensional visualization, the study also reveals demographic and procedural patterns, offering insights into resource allocation, early diagnosis, and risk stratification. This integrated big data approach emphasizes the value of advanced analytics for improving clinical decision-making and personalized medicine in chronic disease management

- **Collins et al. [18]** conduct a systematic review to critically assess the methodologies and reporting practices used in developing risk prediction models for Type 2 diabetes. Reviewing 39 studies and 43 models, they found that many models suffered from significant methodological flaws—such as inadequate sample size, poor treatment of continuous variables, and improper handling of missing data. Key issues included frequent use of univariate screening for predictor selection, categorization of continuous variables, and lack of external validation. Less than half of the studies reported how missing data were addressed, and many employed complete case analysis, which risks bias and data loss. Only a minority incorporated advanced methods such as multiple imputation or considered non-linear relationships. Moreover, the use of automated variable selection techniques (e.g., stepwise regression) was common, despite their known limitations in producing unstable models. The authors highlight that such flawed practices compromise the validity and clinical applicability of the models. They advocate for the development of standardized guidelines to improve the rigor and transparency in creating and reporting risk prediction models for diabetes

- **Bhotta et al. [19]** explore the usability of big data analytics applications—specifically artificial intelligence (AI) and machine learning (ML)—in the management of Type 2 Diabetes Mellitus (T2DM), focusing on the healthcare context in India. Their study underscores that while big data applications have been widely adopted in healthcare, there is a significant research gap concerning their usability and the impact of such usability on achieving intended health outcomes. Through a mixed-methods research design, the authors aim to identify factors influencing the usability of AI/ML systems for T2DM management, which they argue is a determinant of system adoption and successful disease management. They emphasize that usability is not synonymous with perceived usefulness, but rather refers to the end-user's ability to effectively interact with the application—affecting efficiency, satisfaction, and clinical decision-making. The paper presents a conceptual model linking characteristics of healthcare big data (volume, velocity, variety, veracity) to software design attributes (e.g., reliability, scalability, security), which in turn influence usability. Their findings suggest that incorporating usability considerations at the architectural design stage can enhance the effectiveness of AI/ML applications, ultimately contributing to better T2DM outcomes and general improvements in chronic disease management

- **Khanra et al. [20]** conducted a comprehensive systematic literature review to explore the applications and implications of Big Data Analytics (BDA) in healthcare. They synthesized findings from 41 high-quality studies published up to June 2019 and organized the applications of BDA into five primary contexts: health awareness, stakeholder interaction within the healthcare ecosystem, hospital management practices, treatment of specific medical conditions, and technological integration in service delivery. The study found that BDA enables enhanced disease prediction, decision support, personalized patient care, and strategic hospital management by leveraging structured and unstructured data from diverse sources such as electronic health records and IoT-enabled sensor devices. Furthermore, the authors proposed a six-component framework highlighting the interconnected roles of medical records, sensor data, ethical considerations, technology integration, hospital management, and customized patient care. The review emphasized that while BDA holds transformative potential for healthcare efficiency and patient outcomes, it also faces critical challenges related to data quality, privacy, and implementation costs. Khanra et al. conclude by suggesting a future research agenda that includes conceptual advancements, methodological rigor, and explorations of emerging technologies like cloud computing and artificial intelligence within healthcare BDA applications

- **Macinati and Anessi-Pessina et al. [21]** conducted an empirical study to examine the impact of management accounting (MA) use on financial performance within public healthcare organizations in Italy, employing a contingency theory framework. Their model assessed how strategy, organizational size, and regional policy influenced MA design, which in turn affected satisfaction with MA and its actual use by managers. Using structural equation modeling on data from 131 public healthcare institutions, the study found that cost-containment strategies significantly drove the adoption of more sophisticated MA systems. Furthermore, MA use was shown to be influenced both directly by MA design and indirectly through top management's satisfaction with it. While the relationship between MA use and financial performance was modest, the results suggested that it is not merely the design but the effective use of MA systems that contributes to improved outcomes. These findings underscore the importance of contextual and behavioral factors in successfully implementing MA in public settings and challenge assumptions that private-sector tools inherently improve performance in public health institutions. The study offers valuable insights for both managers aiming to leverage MA for better decision-making and policymakers considering NPM-driven reforms in healthcare

- **Kamiran and Calders et al. [22]** address the pressing issue of bias in data-driven decision-making by presenting a comprehensive study on discrimination-aware classification. They explore the challenge of learning classifiers that maintain high accuracy while ensuring non-discriminatory outcomes with respect to sensitive attributes such as gender or ethnicity. The authors formalize the trade-off between accuracy and fairness in binary classification tasks and propose a set of data preprocessing techniques to mitigate discrimination prior to model training. These include suppression (removal of sensitive and correlated attributes), massaging (modifying class labels), reweighing (assigning weights to instances), and sampling (resampling to balance discrimination). Notably, they demonstrate through both theoretical analysis and empirical experiments that naive removal of sensitive attributes does not eliminate discrimination—a phenomenon they call redlining. Their findings show that methods like massaging and preferential sampling significantly reduce discrimination with minimal loss in accuracy. The techniques were validated on real-world datasets including UCI's Census Income data, and the results highlighted the need for ethically aware model development processes in domains subject to anti-discrimination laws

- **Perveen et al. [23]** present a comparative performance analysis of data mining classification techniques for predicting diabetes mellitus using real-world clinical data. Their study utilizes the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database and focuses on identifying effective machine learning models for accurate classification of diabetes across three adult age groups: 18–35, 36–55, and over 55. The authors evaluate three classification strategies: standalone J48 decision tree, bagging, and Adaboost ensembles (with J48 as the base learner). Data preprocessing included selection of key risk factors such as BMI, blood glucose, lipid profiles, and blood pressure measurements. Through experimentation involving 4,678 patients, the results showed that Adaboost outperformed both bagging and the standalone J48 in terms of discriminative capability, especially for smaller sample sizes. Meanwhile, bagging excelled with larger datasets. The performance was measured using the Area Under the Receiver Operating Characteristic Curve (AUROC), and Adaboost consistently showed higher reliability. The study concludes that ensemble methods, particularly Adaboost, offer more robust prediction frameworks for chronic disease classification tasks and suggests extending this methodology to other diseases like hypertension or coronary heart disease.

- **Shyni et al. [24]** present an extensive review of the applications, issues, and challenges associated with implementing big data analytics in the diagnosis of diabetes mellitus. The study emphasizes the transformative potential of big data in healthcare, especially through predictive analytics, prescriptive modeling, and real-time decision-making capabilities. The authors explain that by leveraging massive structured and unstructured datasets—from clinical records to real-time sensor data—healthcare professionals can identify at-risk diabetic patients early and reduce costs and readmission rates. However, they also underline significant challenges including data complexity, storage limitations, security concerns, processing time, and the cost of analytics infrastructure. The paper surveys various predictive and diagnostic applications using technologies like Hadoop, MapReduce, and cloud platforms. Examples include predictive models for diabetes using EMR data, fraud detection in health insurance, retinal image analysis for diabetic retinopathy, and mobile health monitoring systems. The study highlights that while big data analytics shows great promise in improving diagnostic accuracy and treatment outcomes for diabetes, scalability and privacy remain substantial barriers to broader adoption in clinical practice

- **Nagarajan and Chandrasekaran et al. [25]** developed an expert clinical system to diagnose diabetes types and severity using data mining techniques. They applied the Simple K-Means algorithm to classify diabetes into type-1, type-2, and gestational diabetes, followed by four classification algorithms to assess risk levels. SimpleCart achieved perfect classification accuracy. The study identified key attributes, such as HbA1c and plasma glucose, in predicting diabetes risk. The authors suggest that combining clustering and classification can improve early diagnosis and risk assessment, with potential for broader healthcare applications.

The reviewed literature highlights the extensive role of Big Data Analytics (BDA) in healthcare, focusing on predictive analytics, decision support, and real-time monitoring. Several studies emphasize the integration of IoT, AI, and ML for optimizing healthcare operations, though challenges such as data privacy, security, and interoperability remain (Karatas et al. [1]; Palanisamy et al. [2]; Ahmed et al. [8]). Research on diabetes prediction explores ML models like XGBoost, SVM, and ensemble techniques, but limitations in data preprocessing and model validation persist (Hasan et al. [7]; George et al. [15]; Perveen et al. [23]). Studies highlight the role of Hadoop and risk prediction models, yet face challenges in handling unstructured data and standardization (Eswari et al. [16]; Collins et al. [18]). Bhotta et al. [19] and Nazir et al. [3] address Big Data applications in diabetes management, stressing predictive analytics but lacking usability optimizations. While blockchain and AI-driven analytics offer potential solutions, real-time implementations remain underdeveloped (Dash et al. [11]; Wang et al. [12]). Our paper advances these studies by integrating advanced ML techniques with robust data preprocessing, ensuring high accuracy while addressing interoperability and scalability issues. It also enhances predictive modeling through improved validation strategies, refining diabetes diagnosis and management.

3. PROPOSED SYSTEM

3.1. Methodology

This research introduces a strong and holistic method of diabetes status prediction based on the Diabetes_012_Health_Indicators_BRFSS2021 dataset. The approach combines the use of classical machine learning classifiers and deep learning to improve model strength and prediction accuracy.

First, the data is preprocessed by verifying if there are missing values and then separating the feature set (X) from the target variable (Diabetes_012). To have balanced representation across the three diabetes classes (0: No Diabetes, 1: Pre-Diabetes, 2: Diabetes), a stratified train-test split is used.

We train and test several machine learning classifiers such as Decision Tree, Random Forest, Gradient Boosting, Naive Bayes, and Logistic Regression. Each of these models is checked based on important metrics like accuracy, confusion matrix, classification report, and ROC-AUC curves for multi-class classification by label binarization.

To investigate nonlinear correlations and enhance predictive accuracy, we employ an Artificial Neural Network (ANN) with TensorFlow Keras. The ANN has an input layer (equivalent to the feature dimension), two hidden layers with ReLU activation, and a softmax output layer for multi-class classification. The model is sparsely_categorical_crossentropy loss compiled and Adam optimized. Early stopping is used to prevent overfitting.

Lastly, comparison is done for all models on accuracy and ROC-AUC metrics. Visual aids such as confusion matrices and ROC curves are utilized to verify and interpret performance. The hybrid method allows for extensive assessment of algorithmic efficacy in managing big healthcare data in diabetes prediction.

3.1.1. Architecture Diagram

The flowchart in the figure illustrates the entire workflow for a machine learning project on diabetes prediction based on the BRFSS dataset. The process starts with data acquisition, i.e., health indicators from the BRFSS dataset. After data acquisition, Exploratory Data Analysis (EDA) is done to comprehend the distribution of the data, identify patterns, and detect any anomalies.

The second step is data preprocessing, which consists of some vital sub-processes: outlier handling, missing value treatment, and feature selection. In this case, the target variable to predict is 'Diabetes_012', which separates people on the basis of their diabetes status. All these preprocessing steps are very important to ensure the quality and integrity of the dataset prior to providing it as an input to machine learning algorithms.

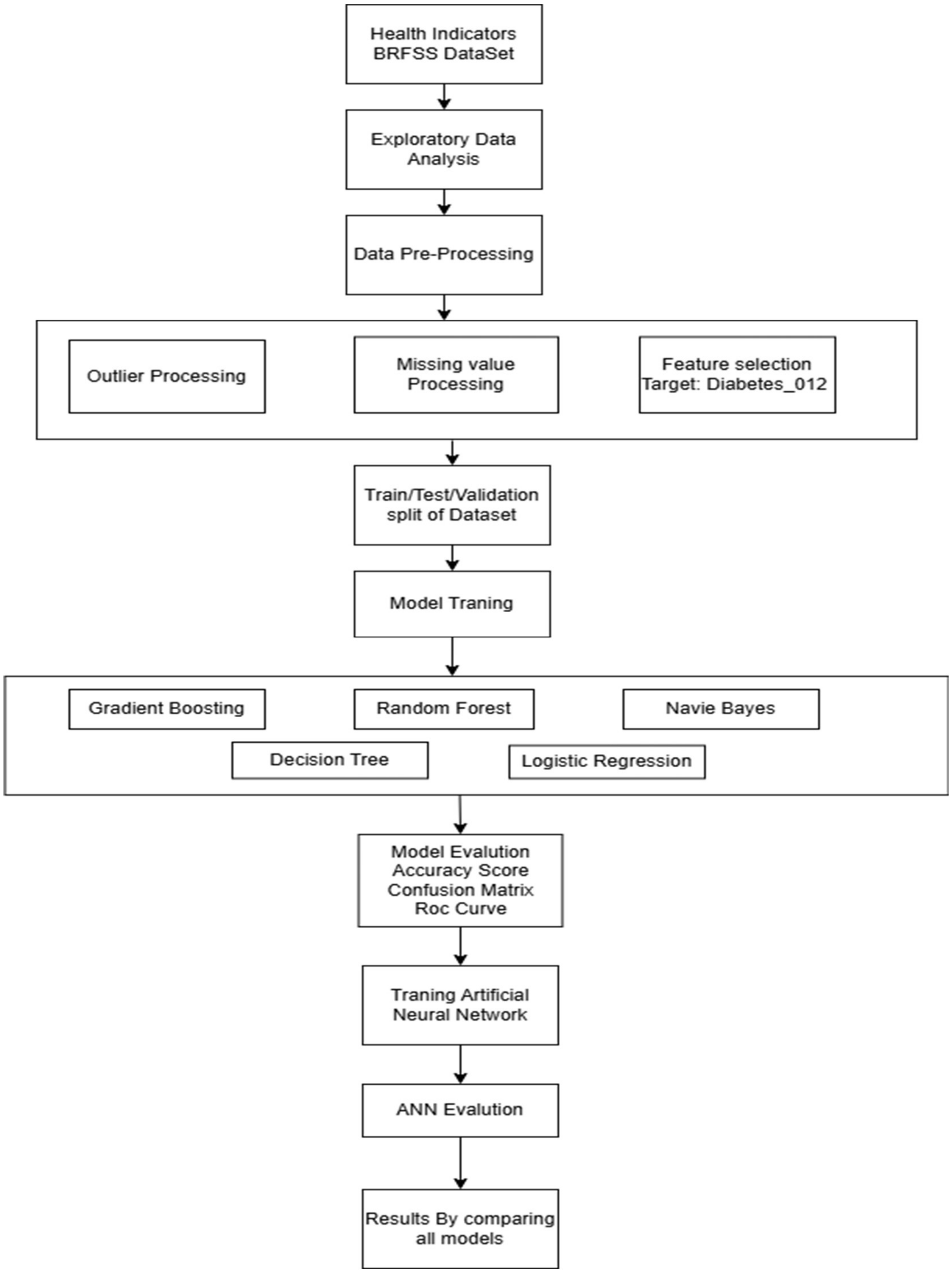

Health Indicators BRFSS DataSet
Exploratory Data Analysis
Data Pre-Processing
Outlier Processing
Missing value Processing
Feature selection Target: Diabetes_012
Train/Test/Validation split of Dataset
Model Traning
Gradient Boosting
Random Forest
Navie Bayes
Decision Tree
Logistic Regression
Model Evalution Accuracy Score Confusion Matrix Roc Curve
Traning Artificial Neural Network
ANN Evalution
Results By comparing all models

Fig 3.1 Architecture Diagram on BDA for diabetics prediction

The dataset is then divided into training, testing, and validation sets to provide a balanced evaluation and avoid data leakage. The training is performed using various classification algorithms like Gradient Boosting, Random Forest, Naive Bayes, Decision Tree, and Logistic Regression. The models are tested with metrics like accuracy score, confusion matrix, and ROC curve to measure their performance and select the top-performing models.

After conventional model training, an Artificial Neural Network (ANN) is also trained on the same data to investigate deep learning potential for the prediction task. ANN testing is done in a similar manner using appropriate performance measures.

The last step is to compare the prediction results from all models, both machine learning models and the ANN, to identify the best-performing model for diabetes prediction. The complete pipeline guarantees a high-quality and reliable model selection process, including how to predict diabetes from health indicators. This systematic approach is crucial in developing a reliable and explainable healthcare prediction system.

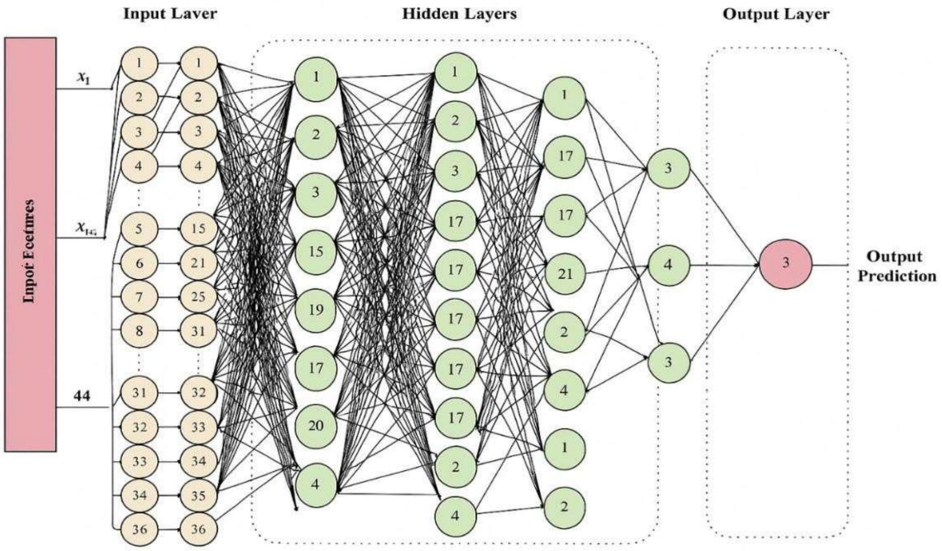


Fig 3.2 Architecture Diagram of ANN

1. Input Layer:

- Receives 44 input features (represented by x_1 to x_{44}), corresponding to the variables from the dataset.
- These can include medical attributes like glucose levels, age, BMI, etc.

2. **Hidden Layers:**

- This ANN includes multiple densely connected hidden layers (with varying neurons in each).
- These layers apply activation functions (like ReLU) to learn complex patterns in the data.
- The intricate web of connections shows how neurons in one layer are linked to all neurons in the next—representing a fully connected neural network.

3. **Output Layer:**

- The final layer provides the output prediction, which is a class label (in this case, “3”).
- This typically goes through a softmax activation function for multi-class classification, or a sigmoid for binary.

3.1.2. Functional Modules

This project is organized into a number of functional modules that together facilitate the prediction of diabetes levels through various machine learning models and an artificial neural network (ANN).

1. **Data Loading and Preprocessing:**
   The data is loaded from a CSV file with pandas. Early preprocessing is checking for missing values and inspecting the structure of the data. The target Diabetes_012 is isolated from features for the training of models.

2. **Separation of Features and Targets:**
   The features (X) are derived by dropping the target column, whereas the labels (y) hold the values of Diabetes_012, which denote various diabetes categories (0: No diabetes, 1: Pre-diabetes, 2: Diabetes).

3. **Train-Test Split:**
   The data set is divided into training and test sets in an 80:20 ratio by train_test_split to allow for unbiased comparison of all models. Reproducibility is achieved with a fixed random state.

4. **Model Training and Evaluation:**
   Several machine learning models are trained, such as Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, and Naive Bayes. Each model is tested using accuracy, classification reports, confusion matrices, and ROC curves for multi-class performance.

5. **Artificial Neural Network (ANN):**
   A deep learning model is constructed with two hidden layers using Keras. The network employs ReLU activation in the hidden layers and softmax in the output layer. The model is trained with early stopping to avoid overfitting and tested on the test set with the same metrics as the standard models.

6. **Model Comparison and Visualization:**
   Each model's performance is presented through confusion matrices and ROC curves. Accuracy scores are compared for identifying the top-performing model for predicting diabetes.This modular framework facilitates easy testing and improvement, guaranteeing strong diabetes prediction across multiple modeling methods.

3.2. Model Description:

1. Decision Tree Classifier:

The Decision Tree Classifier is a supervised learning algorithm used for classification and regression tasks. It works by splitting the dataset into smaller subsets based on specific conditions or feature values, forming a tree-like structure of decisions. Each internal node of the tree represents a decision on a feature, each branch represents the outcome of that decision, and each leaf node represents a class label. The algorithm selects the best feature to split the data using metrics like Gini Impurity or Information Gain. It recursively performs this process until all data points are classified or a stopping condition is met.

In our diabetes prediction project, the Decision Tree model was trained to classify patients into three categories: 0 (No Diabetes), 1 (Prediabetes), and 2 (Diabetes) based on various health indicators. It offered a simple, visual explanation of how different health factors contribute to diabetes classification. Though it can overfit on complex datasets, it gave us a good baseline for interpretability and initial feature importance insights. The model was evaluated using accuracy, confusion matrix, and ROC curves to assess performance across all three classes.

2. Random Forest Classifier:

Random Forest is an ensemble learning technique based on decision trees. Unlike a single decision tree, Random Forest builds multiple trees using subsets of data and features, and combines their predictions through majority voting. This method reduces overfitting and increases accuracy by averaging the predictions of diverse trees, each trained on a bootstrapped (random) sample. The algorithm introduces randomness both in data sampling and feature selection, which ensures low correlation between individual trees, enhancing model stability and generalization.

In our diabetes prediction project, we used the Random Forest Classifier to improve the robustness and accuracy of predictions. Given the high-dimensional and possibly noisy health data, Random Forest performed well by capturing various patterns without overfitting. It effectively handled categorical and continuous variables and helped identify the most influential features related to diabetes, such as BMI, physical activity, or blood pressure. The model's performance was measured using classification metrics, and it outperformed simpler models in terms of accuracy and ROC AUC scores. Its ability to handle large datasets with higher accuracy made it a strong candidate for healthcare prediction tasks like ours.

3. Gradient Boosting Classifier:

Gradient Boosting is a powerful ensemble machine learning algorithm that builds models sequentially. Each new model attempts to correct the errors made by the previous one by focusing more on misclassified examples. Unlike Random Forest, which builds trees in parallel, Gradient Boosting builds them one after the other, using gradients (from loss functions) to minimize errors. It often outperforms other models in accuracy, especially on structured or tabular datasets, but can be prone to overfitting without proper tuning.

In this project, Gradient Boosting was employed to predict diabetes risk levels based on individual health indicators. We selected it because of its ability to learn complex, non-linear relationships and prioritize difficult-to-classify samples. It performed exceptionally well in identifying borderline cases (such as prediabetes), improving overall classification balance. With hyperparameter tuning (e.g., learning rate, number of estimators), Gradient Boosting became one of the top-performing models. We visualized its performance through confusion matrices and ROC curves, and it consistently provided high AUC values across all classes. Its effectiveness in learning from mistakes made it a valuable tool in enhancing our model ensemble for accurate diabetes prediction.

4. Logistic Regression:

Logistic Regression is a statistical model used for binary and multiclass classification problems. It estimates the probability of a class label using a logistic (sigmoid or softmax) function. The model finds a linear decision boundary between classes by learning weights for each feature. Though it assumes a linear relationship between input features and the log odds of the output, it is widely used due to its simplicity, interpretability, and efficiency.

In our diabetes prediction project, we applied Logistic Regression as a baseline model to classify patients into three diabetes categories. It was particularly useful for benchmarking because of its straightforward nature and quick training time. While it does not model complex relationships like tree-based or neural models, it gave reasonably good performance with minimal computational cost. It allowed us to interpret the influence of each feature (such as BMI or smoking status) on diabetes risk, thanks to the coefficient outputs. Its multiclass extension using softmax allowed us to handle all three diabetes categories. Though its performance was slightly lower than ensemble models, it served as a reliable comparison point for evaluating the strength of more advanced classifiers.

5. Naive Bayes:

Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem. It assumes that all features are independent given the target class, an assumption that simplifies computation. Despite this “naive” assumption, it performs surprisingly well in many applications, especially those involving high-dimensional data. It calculates the posterior probability of each class based on prior probabilities and likelihoods derived from the training data.

In our project, we used the Gaussian Naive Bayes variant, suitable for continuous features (like BMI or age). It classified the data into three diabetes categories by estimating the probability distribution of each feature assuming a Gaussian distribution. Although the independence assumption is rarely true in real-world health data (as features like BMI and cholesterol may be correlated), Naive Bayes provided a fast and reasonably accurate model. It was especially useful for understanding baseline performance under simplifying assumptions. We assessed its performance using accuracy, confusion matrices, and ROC curves. Its speed and simplicity made it ideal for initial exploration, and while it didn’t outperform the ensemble methods, it offered useful comparative insights for our diabetes prediction model stack.

6. Artificial Neural Network (ANN):

Artificial Neural Networks (ANNs) are a class of deep learning models inspired by the structure and function of the human brain. An ANN consists of layers of interconnected nodes (neurons), including an input layer, hidden layers, and an output layer. Each neuron receives inputs, applies weights, adds a bias, and passes the result through an activation function. This allows the network to model complex, non-linear relationships.

In our diabetes prediction project, we built an ANN using TensorFlow and Keras. The model had an input layer matching the number of health features, followed by two hidden layers with 64 and 32 neurons (using ReLU activation), and a final softmax output layer for multiclass classification. We trained the model using the Adam optimizer and sparse categorical cross-entropy loss, with early stopping to prevent overfitting. The ANN performed exceptionally well in capturing subtle patterns in the data and provided competitive accuracy compared to tree-based models. Its flexibility allowed it to model non-linear relationships among features like BMI, physical activity, and heart health indicators. The model was evaluated using classification metrics, confusion matrices, and ROC curves. ANN added depth to our model suite by leveraging its ability to learn from complex feature interactions, making it a powerful component of our diabetes prediction pipeline.

3.2.1. Dataset Description

The dataset file, ‘diabetes_012_health_indicators_BRFSS2021.csv’, originates from the 2021 Behavioral Risk Factor Surveillance System (BRFSS) and is designed to support the prediction and analysis of diabetes-related health outcomes. It comprises 236,378 entries and includes 22 health indicators covering a broad spectrum of behavioral, demographic, and clinical features. The target variable, `Diabetes_012`, classifies individuals into one of three categories: 0 for no diabetes, 1 for pre-diabetes, and 2 for diabetes. The dataset features both categorical and continuous variables, including key health attributes such as body mass index (BMI), physical activity, fruit and vegetable consumption, smoking and alcohol habits, cholesterol and blood pressure status, and access to healthcare. Socioeconomic factors like income, education, and age group are also represented. This comprehensive and clean dataset is ideal for building predictive models and gaining insights into the factors contributing to diabetes risk in the U.S. adult population.

Dataset Overview:

- **Shape** **:** 236378 rows x 22 columns
- **Data Types :**
  - 13 Columns are float64
  - 9 columns are int64

S.No	Column Name	Type	Description
1	Diabetes_012	float64	Target variable: 0 = No diabetes, 1 = Pre-diabetes, 2 = Diabetes
2	HighBP	Int64	High Blood Pressure (0 = No, 1 = Yes)
3	HighChol	float64	High Cholesterol (0 = No, 1 = Yes)
4	CholCheck	Int64	Cholesterol check in past 5 years (0 = No, 1 = Yes)

5	BMI	float64	Body Mass Index
6	Smoker	float64	Smoked at least 100 cigarettes in life (0 = No, 1 = Yes)
7	Stroke	float64	Ever had a stroke (0 = No, 1 = Yes)
8	HeartDiseaseAttack	float64	Coronary Heart Disease or Myocardial Infarction (0 = No, 1 = Yes)
9	PhysActivity	Int64	Physical activity in past 30 days (0 = No, 1 = Yes)
10	Fruits	Int64	Consumes fruits 1+ times per day (0 = No, 1 = Yes)
11	Veggies	Int64	Consumes vegetables 1+ times per day (0 = No, 1 = Yes)
12	HvyAlcoholConsump	Int64	Heavy alcohol consumption (0 = No, 1 = Yes)
13	AnyHealthcare	Int64	Has any form of health coverage (0 = No, 1 = Yes)
14	NoDocbcCost	float64	Couldn’t see doctor due to cost in past year (0 = No, 1 = Yes)
15	GenHlth	float64	General Health (1 = Excellent to 5 = Poor)
16	MenHlth	float64	Number of mentally unhealthy days in past 30 days (0–30)
17	PhyHlth	float64	Number of physically unhealthy days in past 30 days (0–30)
18	DiffWalk	float64	Serious difficulty walking/climbing stairs (0 = No, 1 = Yes)

19	Sex	Int64	0 = Female, 1 = Male
20	Age	Int64	Age category (1–13): 13 age brackets from 18 to 80+
21	Education	float64	Education level (1 = Never attended, 6 = College 4+ years)
22	Income	float64	Income level (1 = < $10k, 11 = ≥ $75k)

Table 3.1 Dataset Description

3.2.1.1. Histogram of Features:

As shown in Figure 3.3, the histograms illustrate the distribution of three binary health-related variables from the dataset:

- HighBP (High Blood Pressure)
    - Indicates whether an individual has been diagnosed with high blood pressure.
    - The histogram shows a class imbalance, with more individuals not having high blood pressure (0) than those who do (1).
- HighChol (High Cholesterol)
    - Reflects whether an individual has high cholesterol levels.
    - Similarly, there are fewer individuals with high cholesterol compared to those without, as evident from the taller bar at 0.
- CholCheck (Cholesterol Check)
    - Represents whether the individual has had their cholesterol checked in the past five years.
    - Unlike the previous two, this feature is heavily skewed toward 1, indicating that the majority of participants have undergone a cholesterol check.

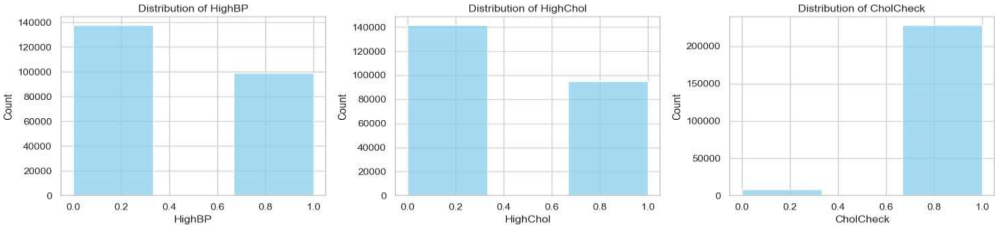


Fig 3.3 Histogram of Features HighBP, HighCol and CholCheck

As shown in Figure 3.4, the histogram provides insight into the distribution of three key health-related features from the dataset:

- BMI (Body Mass Index)
  - A continuous variable measuring body fat based on height and weight.
  - The histogram is right-skewed, with most BMI values clustering between 20 and 40, typical of a general adult population. A few outliers extend toward higher BMI values, indicating obesity in some individuals.
- Smoker
  - A binary categorical variable indicating whether the individual has smoked at least 100 cigarettes in their lifetime.
  - The chart shows a moderate class imbalance, with more non-smokers (0) than smokers (1), which aligns with general health trends in many populations.
- Stroke
  - Another binary variable that indicates whether the person has experienced a stroke.
  - This feature is highly imbalanced, with the vast majority of individuals not having suffered a stroke (0), and only a small proportion marked as having had one (1).

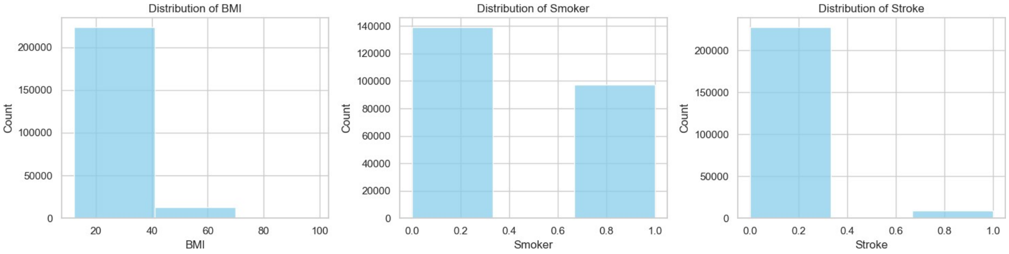


Fig 3.4 Histogram of Features BMI, Smoker and Stroke.

➢ As shown in Figure 3.5, the histogram provides insight into the distribution of three key health-related features from the dataset:

➢ HeartDiseaseorAttack

- This binary variable indicates whether an individual has ever had a heart disease or heart attack.
- The distribution is highly imbalanced, with a significant majority of individuals not having reported heart issues (0), while only a small fraction has (1), reflecting the relatively lower prevalence of such conditions in the general population.

➢ PhysActivity

- This binary feature represents whether an individual has participated in any physical activity in the past month.
- The plot shows that most individuals are physically active (1), which is a positive sign in terms of general health and wellness.

➢ Fruits

- Indicates whether a person consumes fruits at least once a day (1) or not (0).
- The distribution is slightly imbalanced but fairly split, with a higher proportion of people reporting regular fruit intake, suggesting a decent level of dietary awareness among the participants.

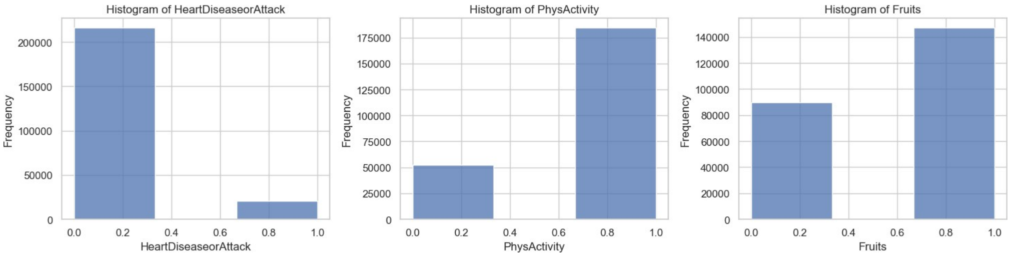


Fig 3.5 Histogram of Features HeartDiseaseorAttack, PhysActivity and Fruits.

➢ As shown in Figure 3.6, the histogram provides insight into the distribution of three key health-related features from the dataset:

➢ Veggies

- Represents whether individuals consume vegetables at least once per day (1) or not (0).

- The majority of individuals reported regular vegetable consumption, indicating relatively healthy dietary habits among most participants.

➢ HvyAlcoholConsump

- Indicates heavy alcohol consumption (yes = 1, no = 0).
- The distribution is highly skewed toward 0, suggesting that most individuals do not engage in heavy drinking, which is a positive health indicator.

➢ AnyHealthcare

- Denotes whether an individual has any form of healthcare coverage (1) or not (0).
- There is a strong imbalance, with the overwhelming majority reporting having healthcare access, highlighting broad coverage in the sample population.

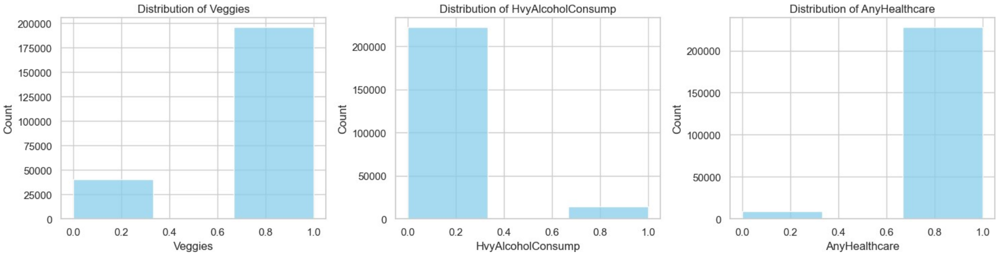


Fig 3.6 Histogram of Features Veggies, HvyAlcoholConsump and AnyHealthcare.

➢ As shown in Figure 3.7, the histogram provides insight into the distribution of three key health-related features from the dataset:

➢ NoDocbcCost

- This binary variable indicates whether individuals were unable to see a doctor due to cost (1 = yes, 0 = no).
- Most individuals reported no financial barrier to accessing healthcare, which reflects positively on affordability and access to medical services.

➢ GenHlth (General Health)

- A categorical variable typically ranging from 1 (excellent) to 5 (poor).
- The distribution is skewed toward better health ratings (1 and 2), suggesting that most respondents perceive their general health to be good to excellent.

- MentHlth (Mental Health Days)
    - Represents the number of days in the past month that mental health was not good (ranging from 0 to 30).
    - The majority of responses cluster around 0, indicating that most individuals reported few to no mentally unhealthy days, though a smaller group reported higher values.

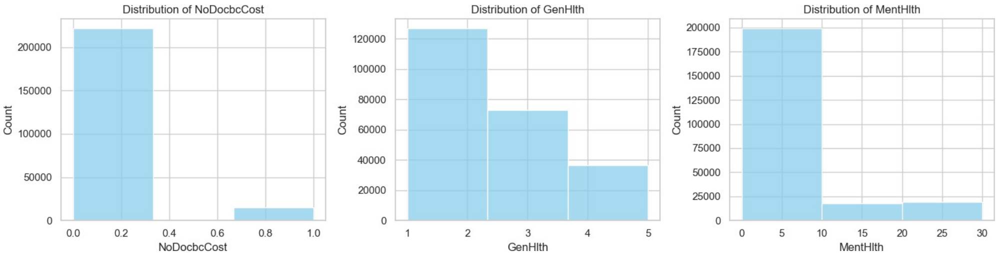


Fig 3.7 Histogram of Features NoDocbcCost, GenHlth and MentHlth

- As shown in Figure 3.8, the histogram provides insight into the distribution of three key health-related features from the dataset:
- PhysHlth (Physical Health Days)
    - Measures the number of days in the past month when physical health was not good (0 to 30).
    - The distribution is heavily skewed toward 0, indicating that most individuals experienced few or no days of poor physical health.
- DiffWalk (Difficulty Walking)
    - A binary indicator of whether the person has serious difficulty walking or climbing stairs (1 = yes, 0 = no).
    - A large portion of the population does not experience mobility issues, but there is still a notable minority that does.
- Sex
    - Coded as binary (0 and 1), representing the sex of the individuals (typically 0 = female, 1 = male or vice versa, depending on dataset documentation).
    - The distribution appears fairly balanced, though there may be a slightly higher number of females (or males), depending on label mapping.

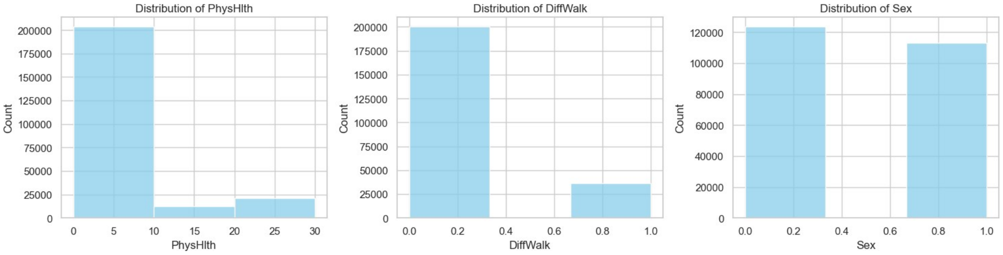


Fig 3.8 Histogram of Features PhysHlth , DiffWalk and Sex

- As shown in Figure 3.9, the histogram provides insight into the distribution of three key health-related features from the dataset:
- Age
  - Coded as a grouped variable (likely 13 levels from younger to older adults).
  - The histogram shows a higher frequency in older age groups, suggesting that the dataset contains more middle-aged and senior individuals, which is common in health-related datasets.
- Education
  - Categorized likely from 1 (lowest education level) to 6 (highest).
  - The majority of the population falls into higher education categories, particularly level 5, indicating that most individuals have completed high school or higher education.
- Income
  - Categorized income levels from 1 (lowest income) to 8 or 9 (highest), often with mid-levels representing ranges.
  - Most individuals fall into the middle to upper income brackets, reflecting a relatively balanced distribution but skewed slightly toward the higher end.

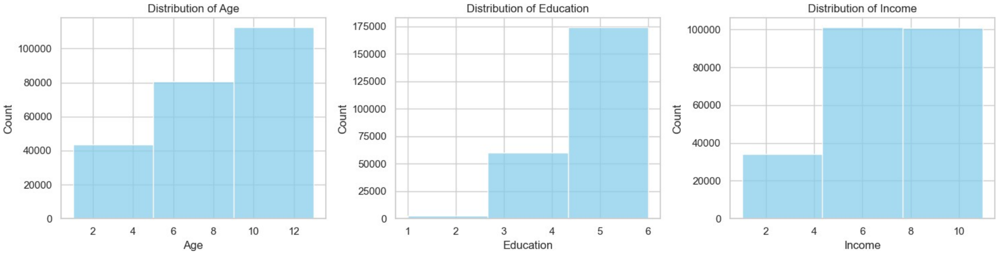


Fig 3.9 Histogram of Features Age , Education and Income.

3.2.1.2. Violin Graphs Comparison

As shown in Figures 3.10, 3.11 and 3.12 the violin plots visually compare the distribution of several health and socioeconomic indicators against diabetes status using the dataset. Each plot reveals how a specific feature varies across the three diabetes classes: 0, 1, and 2.

1. Education vs Diabetes Status

- X-axis: Diabetes status (0, 1, 2)
- Y-axis: Education level (1 = Least educated, 6 = Most educated)
- ➢ Observation:
  - Individuals with diabetes (class 2) are more concentrated in lower education levels compared to non-diabetics.
  - Higher education levels (5–6) are more common among non-diabetics, suggesting an inverse relationship between education and diabetes risk.

2. Income vs Diabetes Status

- X-axis: Diabetes status
- Y-axis: Income level (1 = Lowest, 11 = Highest)
- ➢ Observation:
  - Non-diabetics (class 0) tend to fall in higher income brackets.
  - Diabetic individuals (class 2) show a strong concentration in lower income levels.

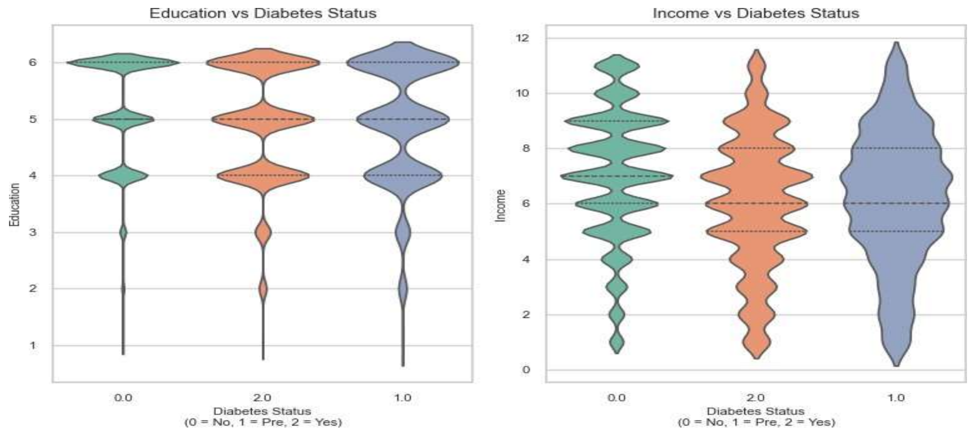


Fig 3.10 Comparison of Diabetes Statues with Education and Income using Violin Plots.

3. **GenHlth (General Health) vs Diabetes Status**

- X-axis: Diabetes status
- Y-axis: General health score (1 = Excellent, 5 = Poor)

➢ Observation:

  - Non-diabetics are skewed towards better health (1–2).
  - Diabetic individuals report poorer general health, clustering around scores 4–5.
  - Indicates that poorer self-rated health correlates with diabetes.

4. **MentHlth (Mentally Unhealthy Days) vs Diabetes Status**

- X-axis: Diabetes status
- Y-axis: Mentally unhealthy days in the past 30 days (0–30)

➢ Observation:

  - Non-diabetics mostly report 0 mentally unhealthy days.
  - Diabetics show a broader spread, with some experiencing up to 30 days.
  - Mental health challenges may be more prevalent among diabetic individuals.

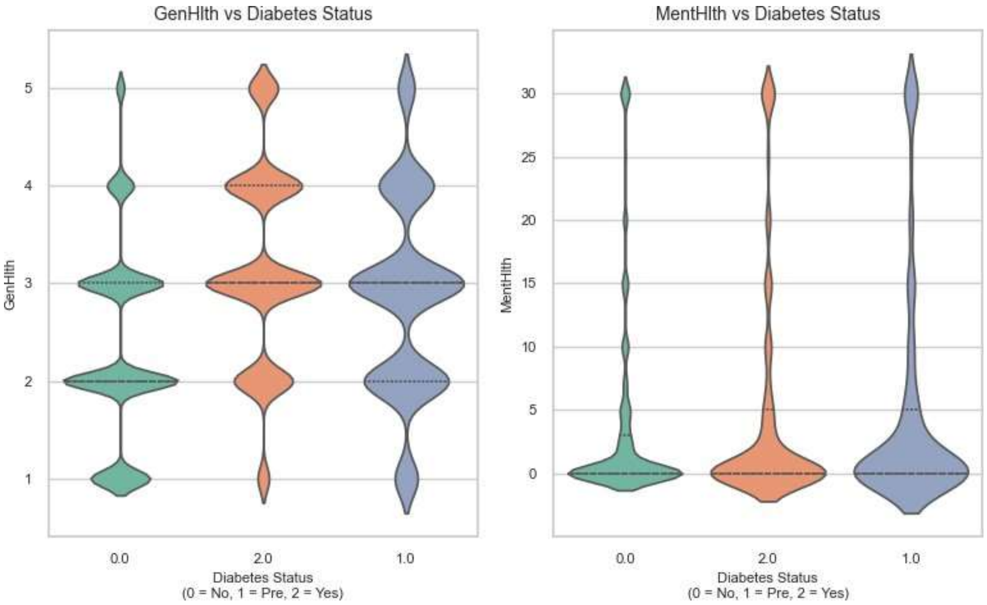


Fig 3.11 Comparison of Diabetes Statues with GenHlth and MentHlth using Violin Plots.

5. PhysHlth (Physically Unhealthy Days) vs Diabetes Status

- X-axis: Diabetes status
- Y-axis: Physically unhealthy days in the past 30 days (0–30)
- Observation:
    - Similar to mental health, non-diabetics mostly report 0–5 days.
    - Diabetics have a much wider distribution, indicating more physically unhealthy days.
    - Strong association between physical health limitations and diabetes.

6. BMI vs Diabetes Status

- X-axis: Diabetes status
- Y-axis: Body Mass Index (BMI)
- Observation:
    - BMI is noticeably higher in individuals with diabetes.
    - The distribution for class 2 (diabetes) shifts right, with many values >30 (obese range).
    - Higher BMI is a strong risk factor for diabetes.

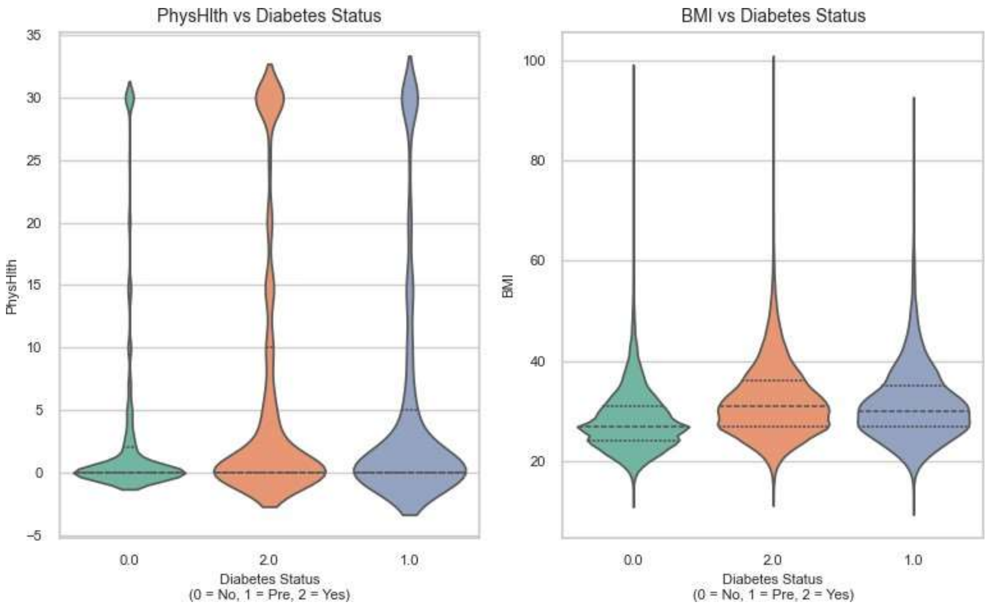


Fig 3.12 Comparison of Diabetes Statues with PhysHlth and BMI using Violin Plots.

4. EXPERIMENTAL SETUP & RESULTS

4.1. System Specifications

The experiments were conducted on a local machine with the following specifications:

- Processor: Intel Core i7 / AMD Ryzen 7 or equivalent
- RAM: 16 GB DDR4
- Storage: 512 GB SSD
- GPU: NVIDIA GeForce GTX/RTX (Optional for training acceleration)
- Operating System: Windows 11 / Ubuntu 20.04

4.1.1. Software Requirements

Software	Version
Python	3.10+
TensorFlow / Keras	2.10+
NumPy	1.24+
Pandas	2.0+
Scikit-learn	1.2+
Matplotlib / Seaborn	For visualizations
Graphviz / Pydot	For model architecture diagrams

Table 4.1. Software Requirements

All dependencies were installed and managed using pip in a virtual environment for isolation.

4.2. Description

This project implements a machine learning and deep learning pipeline for multi-class diabetes prediction using a publicly available health indicators dataset. The pipeline includes:

- Data Preprocessing:
  The dataset is cleaned by checking for missing values and splitting into features and labels. The data is further split into training and testing sets using an 80/20 ratio to prepare for model training.
- Model Training & Evaluation:
  Multiple machine learning models are trained and evaluated, including:
  - Decision Tree
  - Random Forest
  - Gradient Boosting
  - Logistic Regression
  - Naive Bayes

  Each model is evaluated using accuracy, classification report, confusion matrix, and multi-class ROC-AUC curves for a comprehensive performance comparison.
- Artificial Neural Network (ANN):
  A custom ANN model is built using TensorFlow/Keras Sequential API, consisting of:
  - Dense layers (64 → 32 units with ReLU activation)
  - Output layer with softmax activation for 3-class classification

  The model is compiled with the Adam optimizer and sparse categorical cross-entropy loss, trained with early stopping for generalization, and validated with performance metrics and ROC-AUC curves.
- Visualization & Insights:
  Model performance is visualized using confusion matrices and ROC curves to compare classification capabilities across all models.

This hybrid approach enables a robust comparison between traditional machine learning and deep learning techniques for early diabetes detection.

4.3. Results & Test Analysis

Metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,Score = 2.\frac{Precision.Recall}{Precision + Recall}$$

1. Decision Tree:

The Decision Tree classifier is a supervised learning algorithm that splits data into branches based on feature thresholds, ultimately leading to leaf nodes representing class predictions. It is simple to interpret and effective for both classification and regression tasks, as shown in Figure 4.1

- The multi-class confusion matrix illustrates predictions across three distinct classes:
  - 0 – Non-diabetic
  - 1 – Pre-diabetic
  - 2 – Diabetic.
- The binary confusion matrix consolidates pre-diabetic and diabetic into a single “Diabetic” category, contrasting it with “Non-diabetic”.

Decision Trees perform relatively well when class boundaries are distinct. However, they may suffer from overfitting, especially in medical datasets with noise or class imbalance. In this case, the Decision Tree shows solid accuracy for non-diabetic predictions but struggles to distinguish pre-diabetics and diabetics—highlighting the need for pruning or ensemble approaches for better generalization.

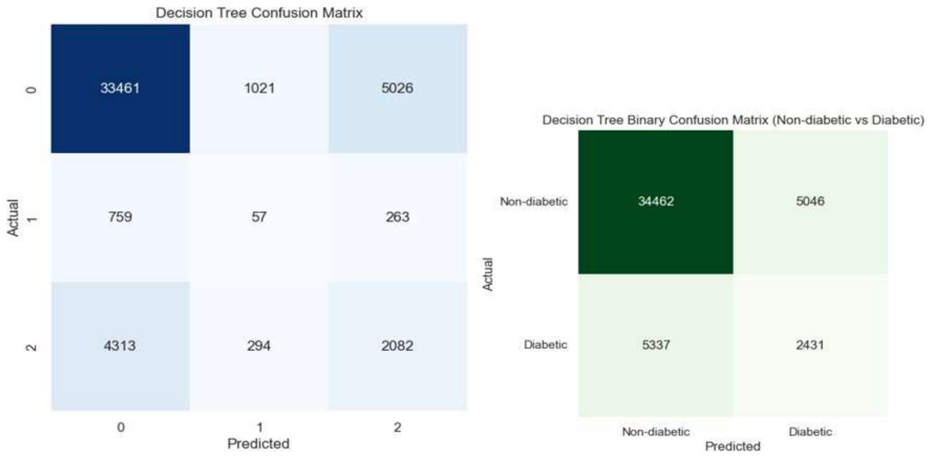


Fig 4.1 Decision Tree Confusion Matrix.

2. Random Forest:

The Random Forest model is an ensemble learning method that combines multiple decision trees to enhance predictive accuracy and control over-fitting. This classifier was applied to a medical dataset for diabetes classification. as shown in Figure 4.2

- The multi-class confusion matrix shows classification across three categories:
  - 0 – Non-diabetic
  - 1 – Pre-diabetic
  - 2 – Diabetic
- The binary confusion matrix simplifies this into:
  - Non-diabetic
  - Diabetic (Pre-diabetic + Diabetic combined)

The model demonstrates good accuracy for non-diabetics, but performance decreases notably for diabetic and pre-diabetic categories, indicating class imbalance or overlapping features.

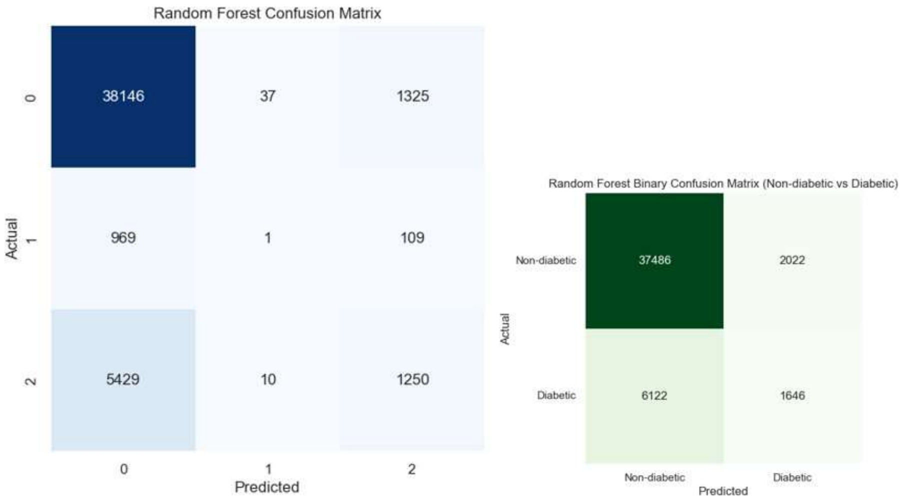


Fig 4.2 Random Forest Confusion Matrix.

3. Gradient Boosting:

Gradient Boosting is a powerful boosting algorithm that builds sequential decision trees where each tree attempts to correct the errors of the previous ones. It's known for handling complex data and non-linear relationships. as shown in Figure 4.3

- The multi-class confusion matrix again targets three classes (0, 1, 2).
- The binary matrix groups predictions into diabetic vs non-diabetic.

This model performs well for class 0 (non-diabetic) but shows very limited correct classification for class 1 (pre-diabetic), suggesting a weakness in handling the middle class.

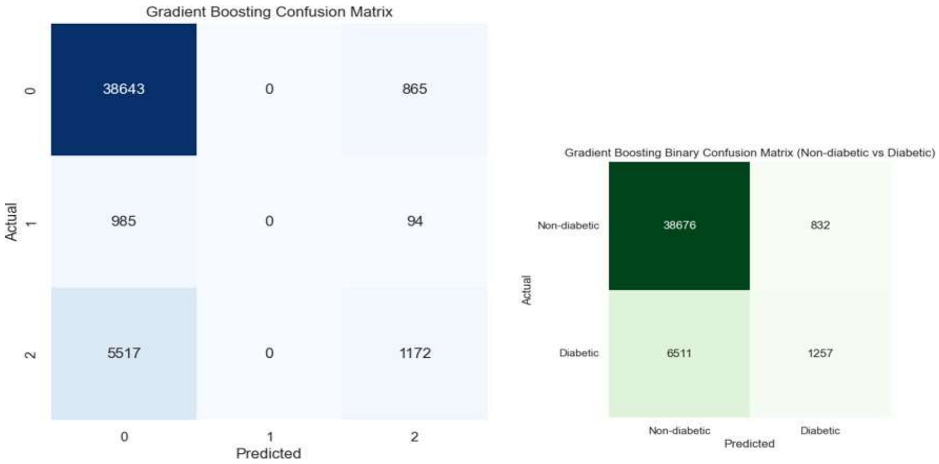


Fig 4.3 Gradient Boosting Confusion Matrix.

4. Logistic Regression:

Logistic Regression is a traditional linear model best suited for binary classification. It was adapted here for multi-class prediction using techniques like one-vs-rest. as shown in Figure 4.4

- The multi-class confusion matrix shows high correct predictions for class 0.
- The binary matrix continues to highlight the same trend: strong performance on non-diabetics but weak recall for diabetics.

This suggests Logistic Regression might not capture non-linear decision boundaries effectively in this context.

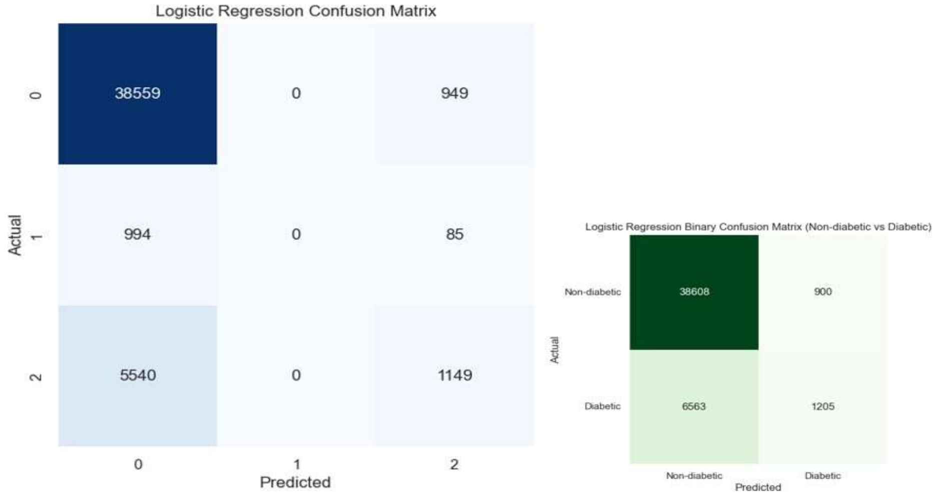


Fig 4.4 Logistic Regression Confusion Matrix.

5. Naive Bayes:

Naive Bayes is a probabilistic classifier based on Bayes' theorem with the “naive” assumption of feature independence. It’s fast and effective for text and categorical data. as shown in Figure 4.5

- The multi-class matrix reveals substantial misclassification across classes, especially between non-diabetics and diabetics.
- In the binary matrix, while diabetic detection is improved compared to logistic regression, many diabetics are still missed.

The model performs quickly but struggles with overlapping data and class nuances, as seen here.

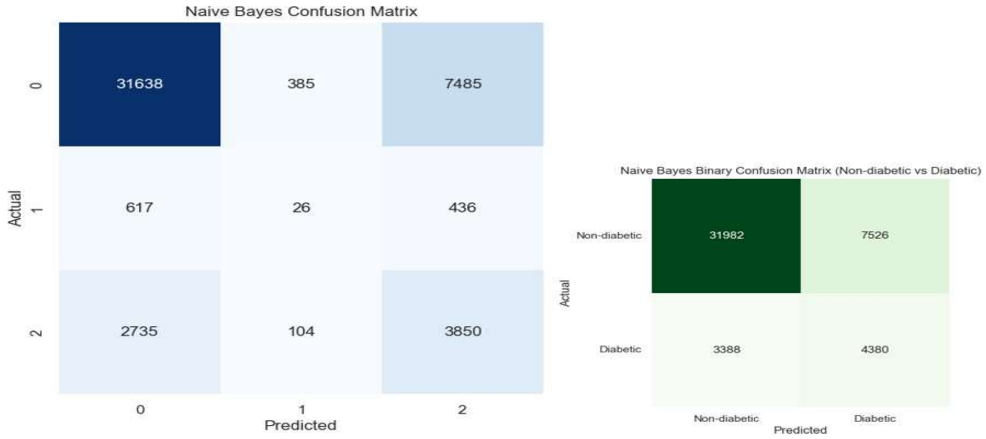


Fig 4.5 Naïve Bayes Confusion Matrix.

6. Artificial Neural Network(ANN):

ANNs are inspired by the human brain and excel in learning complex patterns through multiple layers of neurons. This deep learning model was trained to classify diabetes status. as shown in Figure 4.6

- The multi-class matrix shows excellent prediction for non-diabetics but nearly zero recognition for pre-diabetics.
- In the binary matrix, ANN maintains high accuracy but suffers from low recall for diabetic cases, which is concerning for real-world screening.

ANN demonstrates the capacity to learn well from data.

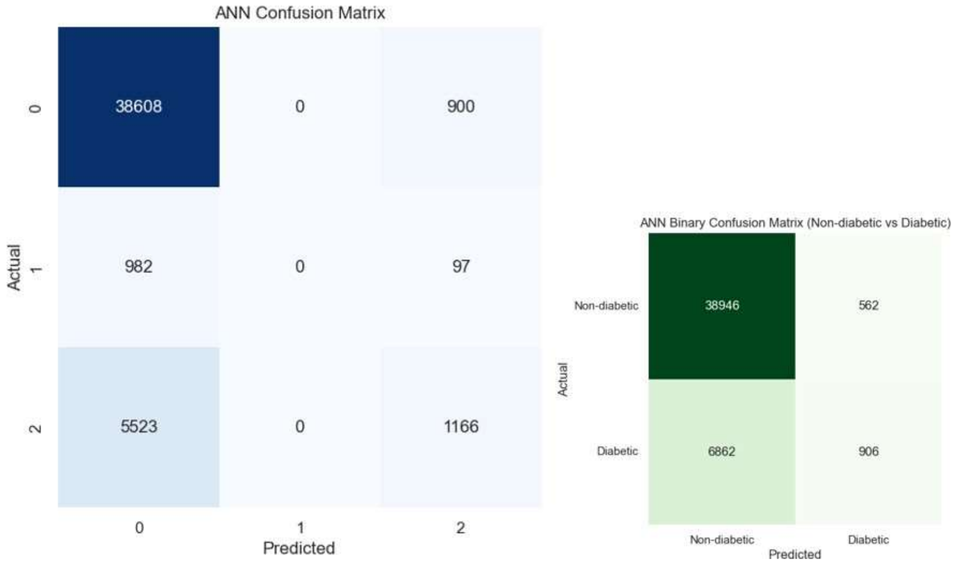


Fig 4.6 ANN Confusion Matrix.

The Figure 4.7 gives a comparative assessment of different classification models by their accuracy scores when used to predict diabetes. Among the models, Gradient Boosting, Artificial Neural Networks (ANN), and Logistic Regression were all able to attain the maximum accuracy of 0.84, which testifies to their high performance and applicability for this classification problem. These models were able to capture the intricate patterns in the data and generalize well on new data.

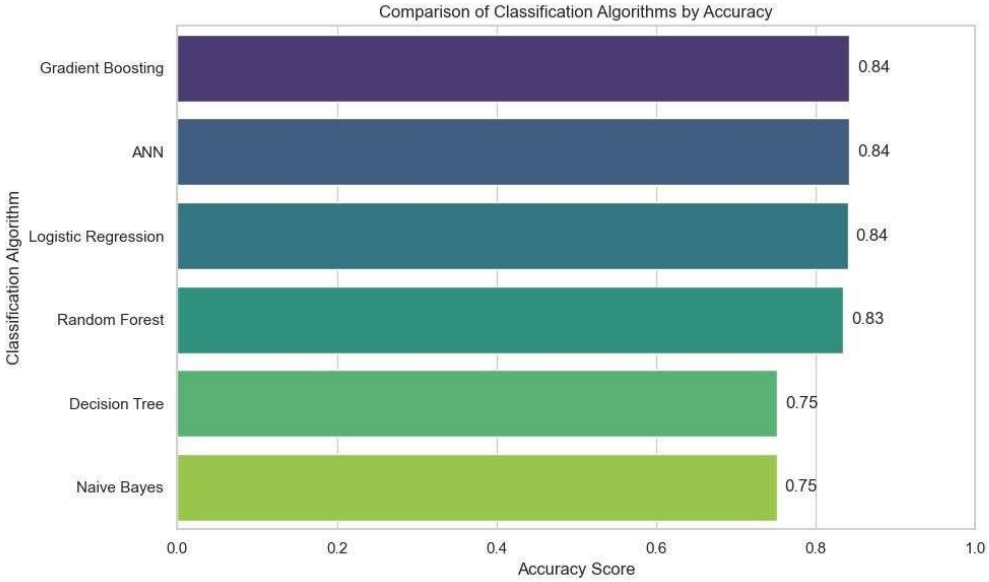


Fig 4.7. Comparision of accuracy with Different Classification Algorithms.

Random Forest also did well with an accuracy of 0.83, and this makes it a strong and stable model because it uses ensemble learning. Decision Tree and Naïve Bayes, although less complex and quicker to train, were behind with an accuracy of 0.75. This indicates that they might not be as good at detecting the subtle relationships in the dataset as more complicated models.

This contrast emphasizes the significance of model selection in obtaining the best predictive performance. It also indicates that ensemble and deep learning techniques offer a considerable benefit in processing structured health data for classification problems. In deployment or further optimization, concentrating on the highest-performing models (Gradient Boosting, ANN, and Logistic Regression) would be the most efficient strategy.

The Figure 4.8 represents the correlation matrix among all numeric features in the dataset, excluding the target variable. Correlation values range from -1 (strong negative) to +1 (strong positive), with values near 0 indicating little to no linear relationship.

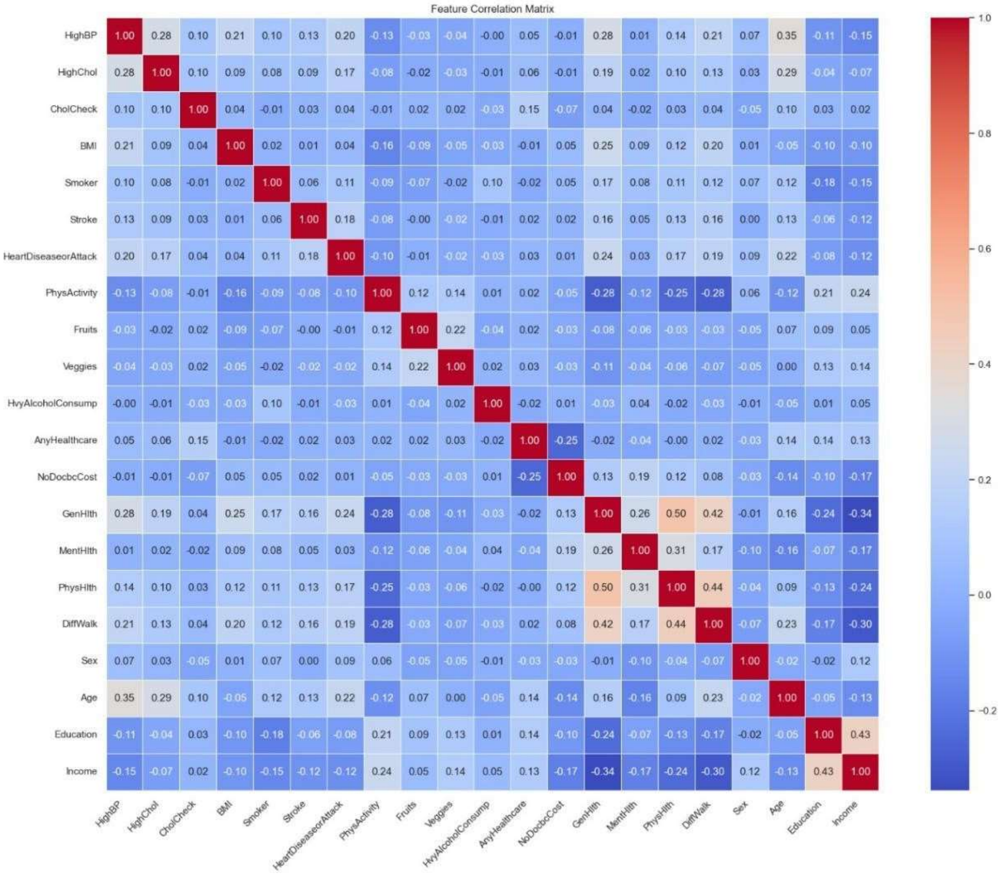


Fig 4.8 Correlation Matrix

The diagonal values are all 1.0, as each feature is perfectly correlated with itself. Among the features, GenHlth shows a strong positive correlation with PhysHlth (0.42) and DiffWalk (0.41), suggesting that individuals with poorer general health tend to report more physically unhealthy days and walking difficulties. Similarly, HighBP and HighChol are moderately correlated (0.28), which aligns with known medical associations.

Negative correlations were observed between PhysActivity and DiffWalk (–0.28), MentHlth and GenHlth (–0.34), and between Income and NoDocbcCost (–0.24), indicating that higher income reduces cost-related barriers to healthcare. Age is positively correlated with HighBP and HighChol, consistent with age-related health risks.

Most feature pairs, however, show weak correlations (values between –0.1 and 0.1), suggesting limited multicollinearity. This ensures that most features provide unique contributions to the model and supports the use of multivariate techniques in prediction.

This correlation analysis is vital for feature selection and understanding variables.

S.No	Model	Target Value	Precision	Recall	F1-Score	Accuracy
1	Decision Tree	0	0.87	0.85	0.86	75%
		1	0.04	0.05	0.05	
		2	0.28	0.31	0.29	
2	Random Forest	0	0.86	0.97	0.91	83%
		1	0.02	0.00	0.00	
		2	0.47	0.19	0.27	
3	Gradient Boosting	0	0.86	0.98	0.91	84%
		1	0.00	0.00	0.00	
		2	0.55	0.18	0.27	
4	Logistic Regression	0	0.86	0.98	0.91	83%
		1	0.00	0.00	0.00	
		2	0.53	0.17	0.26	
5	Naïve Bayes	0	0.90	0.80	0.85	75%
		1	0.05	0.02	0.03	
		2	0.33	0.58	0.42	
6	Artificial Neural Networks	-	-	-	-	84%

Table 4.2 Results

In our results, Gradient Boosting achieved the highest overall accuracy of 84%, outperforming other models such as Decision Trees, Random Forests, Logistic Regression, and Naïve Bayes. Notably, it maintained a high precision and recall for the majority class (class 0), and although its performance on minority classes (1 and 2) was limited—likely due to class imbalance—it still outperformed other models in terms of balanced F1 scores across all classes.

This demonstrates that Gradient Boosting not only captures the majority class well but also shows a relatively better ability to detect minority classes compared to other classifiers, making it the most effective model in our study.

5. CONCLUSION AND FUTURE SCOPE

5.1. Conclusion

The primary objective of this project was to design and implement a comprehensive and efficient system for multi-class diabetes classification , leveraging both classical machine learning (ML) models and deep learning approaches using an Artificial Neural Network (ANN). This project utilized the Behavioral Risk Factor Surveillance System (BRFSS) 2021 dataset, a robust and diverse dataset containing real-world health indicators collected from a large population. This rich dataset provided the foundation for building a predictive model capable of categorizing individuals into three distinct diabetes risk levels: non-diabetic (0) .pre-diabetic (1), and diabetic (2).

The system followed a structured and modular pipeline:

1. Data Preprocessing:
   - Features were analyzed for missing values and appropriately cleaned to ensure high data quality. The target variable (`Diabetes_012`) was separated, and the dataset was split into training and testing sets using a stratified approach to preserve class distribution.

2. Model Training & Evaluation:
   - Several classical machine learning algorithms were employed, including:
     a. Decision Tree Classifier
     b. Random Forest Classifier
     c. Gradient Boosting Classifier
     d. Logistic Regression
     e. Naive Bayes
   - These models were chosen for their diverse strengths—interpretability, ensemble power, generalization, and simplicity. Each model was trained using the training dataset and evaluated using metrics such as:
     a. Accuracy Score for overall performance
     b. Confusion Matrix for per-class prediction analysis
     c. Classification Report for precision, recall, and F1-score

    d. ROC Curves and AUC Scores for a nuanced view of model capability across multi-class classification

- The inclusion of **multi-class ROC analysis using one-vs-rest binarization** allowed for detailed performance insight at the class level, enhancing model comparability.

3. Deep Learning Approach with ANN:
    - An Artificial Neural Network was constructed using the Keras API with TensorFlow backend. The network architecture consisted of:
        a. An input layer aligned with the number of health indicators
        b. Two hidden layers using ReLU activation to capture complex nonlinear relationships
        c. An output layer with softmax activation for multi-class classification
    - To prevent overfitting, early stopping was implemented, which halted training when the validation loss stopped improving, preserving the model's generalization power. The model was evaluated on the testing set using the same metrics as the classical models, and it showed competitive, often superior accuracy—especially in capturing subtle feature interactions across classes.
4. Visualization and Interpretability:
    - The use of Seaborn and Matplotlib for plotting confusion matrices and ROC curves enabled intuitive visual interpretation of each model's strengths and weaknesses. These tools not only helped validate the models but also made the findings more accessible to non-technical stakeholders, such as medical professionals or decision-makers.

Overall, this project demonstrated the practical viability of machine learning and deep learning techniques in medical diagnostics, specifically in predicting diabetes risk levels. The comparative analysis across models highlighted the trade-offs between interpretability and performance. While models like Decision Trees and Logistic Regression offer easier interpretation, more complex models like Random Forests, Gradient Boosting, and ANN provided higher predictive accuracy.

The structured and extensible pipeline developed during this project serves as a strong foundation for future enhancements, including real-time prediction systems, mobile health apps, or integration with electronic health records (EHRs). This work not only satisfies its immediate goal of diabetes classification but also sets the stage for deploying intelligent, data-driven solutions in public health and preventive care domains.

5.2.Future Scope

The most promising aspects of future development are as follows:

- The current system lays a solid foundation for intelligent healthcare diagnostics. To increase the system's usability, accuracy, and real-world applicability, the following future enhancements are proposed:
    - Integration with Electronic Health Records (EHR):
        - The system can be extended to automatically pull patient data from EHRs in hospitals and clinics, providing doctors with real-time diabetes risk assessment.
    - Real-time Mobile App for Self-Screening:
        - A lightweight version of the model can be integrated into a mobile application, allowing users to input lifestyle and health metrics to get instant diabetes risk levels.
    - Explainable AI (XAI) Integration:
        - Integrating tools like SHAP or LIME to provide interpretability to healthcare professionals by explaining how individual features affect the prediction outcome.
    - Time Series Prediction and Risk Progression Modeling:
        - Incorporating temporal data and LSTM (Long Short-Term Memory) networks can allow for modeling progression from pre-diabetic to diabetic states over time, enabling proactive interventions.
    - Hyperparameter Tuning and Automated ML (AutoML):

    - Grid Search and Random Search can be expanded, or frameworks like TPOT or AutoKeras can be implemented to automatically find the most optimal models.

- Feature Engineering and Dimensionality Reduction:
    - Apply advanced feature selection or PCA to improve computational efficiency while retaining key predictive power.

- Cloud Deployment and Scalable APIs:
    - Deploying the trained models via RESTful APIs on cloud platforms (like AWS, GCP, or Azure) for integration into third-party healthcare apps or remote clinics.

- Federated Learning for Privacy-Aware Training:
    - To address data privacy concerns, especially in medical contexts, federated learning can be adopted where model training happens across decentralized devices or institutions without sharing sensitive patient data.

- Multi-modal Diagnosis Systems:
    - Combine numerical health indicator data with other input types like medical imaging, genetic profiles, or wearable sensor data to build a more holistic and accurate diagnostic system.

REFERENCES

[1] Karatas, Mumtaz, et al. "Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives." Expert Systems with Applications 200 (2022): 116912.

[2] Palanisamy, Venketesh, and Ramkumar Thirunavukarasu. "Implications of big data analytics in developing healthcare frameworks–A review." Journal of King Saud University-Computer and Information Sciences 31.4 (2019): 415-425.

[3] Nazir, Shah, et al. "A comprehensive analysis of healthcare big data management, analytics and scientific programming." IEEE Access 8 (2020): 95714-95733.

[4] Wang, Gang, et al. "Big data analytics in logistics and supply chain management: Certain investigations for research and applications." International journal of production economics 176 (2016): 98-110.

[5] Wang, Yichuan, LeeAnn Kung, and Terry Anthony Byrd. "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations." Technological forecasting and social change 126 (2018): 3-13.

[6] Bebortta, Sujit, et al. "DeepMist: Toward deep learning assisted mist computing framework for managing healthcare big data." IEEE Access 11 (2023): 42485-42496.

[7] Hasan, Md Kamrul, et al. "Diabetes prediction using ensembling of different machine learning classifiers." IEEE Access 8 (2020): 76516-76531.

[8] Ahmed, Awais, et al. "Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts." IEEE Access 11 (2023): 112891-112928.

[9] Hussain, Fatima, et al. "Leveraging big data analytics for enhanced clinical decision-making in healthcare." IEEE Access 11 (2023): 127817-127836.

[10] Aceto, Giuseppe, Valerio Persico, and Antonio Pescapé. "The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges." Journal of Network and Computer Applications 107 (2018): 125-154.

[11] Dash, Sabyasachi, et al. "Big data in healthcare: management, analysis and

future prospects." Journal of big data 6.1 (2019): 1-25.

[12] Wang, May D. "Biomedical big data analytics for patient-centric and outcome-driven precision health." 2015 IEEE 39th Annual Computer Software and Applications Conference. Vol. 3. IEEE, 2015.

[13] Raghupathi, Wullianallur, and Viju Raghupathi. "Big data analytics in healthcare: promise and potential." Health information science and systems 2 (2014): 1-10.

[14] Kankanhalli, Atreyi, et al. "Big data and analytics in healthcare: Introduction to the special section." Information Systems Frontiers 18 (2016): 233-235.

[15] George, AS Hovan, et al. "A Survey study on big data analytics to predict diabetes diseases using supervised classification methods." Partners Universal International Innovation Journal 1.1 (2023): 1-8.

[16] Eswari, T., P. Sampath, and S. J. P. C. S. Lavanya. "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50 (2015): 203-208.

[17] Kolesnichenko, Olga, et al. "Big data analytics of inpatients flow with diabetes mellitus type 1: Revealing new awareness with advanced visualization of medical information system data." 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2019.

[18] Collins, Gary S., et al. "Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting." BMC medicine 9 (2011): 1-14.

[19] Bhotta, Dinakar, et al. "An investigation into usability of big data analytics in the management of Type 2 Diabetes Mellitus." 24th Annual Conference of the Asia Pacific Decision Sciences Institute: Full papers. University of Southern Queensland, 2019.

[20] Khanra, Sayantan, et al. "Big data analytics in healthcare: a systematic literature review." Enterprise Information Systems 14.7 (2020): 878-912.

[21] Macinati, Manuela S., and Eugenio Anessi-Pessina. "Management accounting use and financial performance in public health-care organisations: Evidence from the Italian National Health Service." Health Policy 117.1 (2014): 98-111.

[22] Kamiran, Faisal, and Toon Calders. "Data preprocessing techniques for classification without discrimination." Knowledge and information systems 33.1 (2012): 1-33.

[23] Perveen, Sajida, et al. "Performance analysis of data mining classification techniques to predict diabetes." Procedia Computer Science 82 (2016): 115-121.

[24] Shyni, S., R. Shantha Mary Joshitta, and L. Arockiam. "Applications of big data analytics for diagnosing diabetic mellitus: issues and challenges." International Journal of Recent Trends in Engineering & Research 2.06 (2016): 454-461.

[25] Nagarajan, Srideivanai, and R. M. Chandrasekaran. "Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques." Indian Journal of science and Technology 8.8 (2015): 771-6.

APPENDIX

Code:

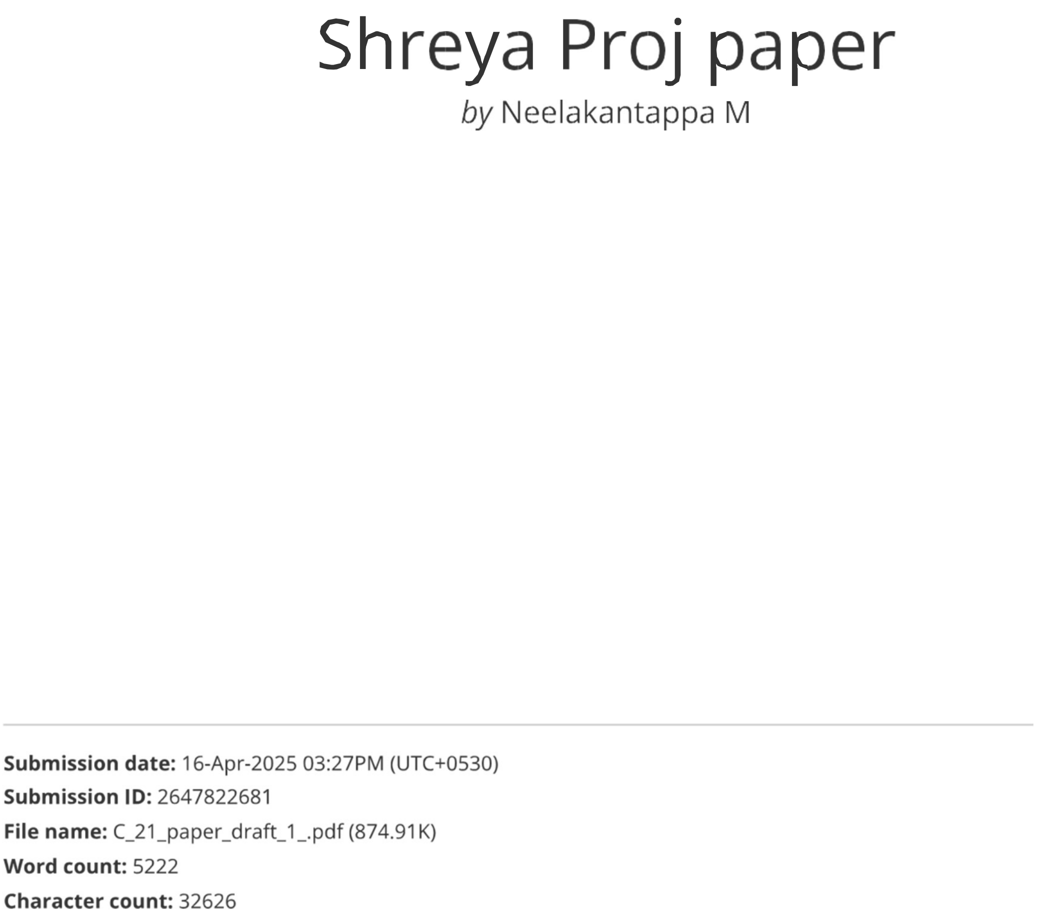
```
# Train and evaluate multiple models
models = {
    'Decision Tree': DecisionTreeClassifier(),
    'Random Forest': RandomForestClassifier(),
    'Gradient Boosting': GradientBoostingClassifier(),
    'Logistic Regression': LogisticRegression(max_iter=1000),
    #'Linear SVM': SVC(kernel='linear'),
    'Naive Bayes': GaussianNB(),
}
accuracy_scores = {}
for name, model in models.items():
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    accuracy_scores[name] = accuracy
    print(f"{name} Accuracy: {accuracy:.4f}")
    print(classification_report(y_test, y_pred))
```

```
# Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', cbar=False)
plt.title(f'{name} Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
# ROC Curve for multi-class
y_test_bin = label_binarize(y_test, classes=[0, 1, 2])
y_pred_bin = label_binarize(y_pred, classes=[0, 1, 2])
fpr = dict()
tpr = dict()
roc_auc = dict()
for i in range(3):
    fpr[i], tpr[i], _ = roc_curve(y_test_bin[:, i], y_pred_bin[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
plt.figure(figsize=(8, 6))
for i in range(3):
    plt.plot(fpr[i], tpr[i], label=f'Class {i} (AUC = {roc_auc[i]:.2f})')
    plt.plot([0, 1], [0, 1], 'k--')
    plt.xlabel('False Positive Rate')
    plt.ylabel('True Positive Rate')
    plt.title(f'{name} ROC Curve')
    plt.legend(loc='lower right')
    plt.show()
    print("="*50)
# Display accuracy scores
accuracy_scores
# Artificial Neural Network (ANN)
model = Sequential()
model.add(Dense(64, input_dim=X_train.shape[1], activation='relu'))
model.add(Dense(32, activation='relu'))
model.add(Dense(3, activation='softmax'))
model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',
metrics=['accuracy'])
early_stopping = EarlyStopping(monitor='val_loss', patience=5,
restore_best_weights=True)
history = model.fit(X_train, y_train, validation_split=0.2, epochs=50, batch_size=32,
callbacks=[early_stopping], verbose=1)
```

```
# Evaluate ANN

y_pred_ann = model.predict(X_test)

y_pred_ann = np.argmax(y_pred_ann, axis=1)

accuracy_ann = accuracy_score(y_test, y_pred_ann)

accuracy_scores['ANN'] = accuracy_ann

print(f"ANN Accuracy: {accuracy_ann:.4f}")

print(classification_report(y_test, y_pred_ann))

# Confusion Matrix for ANN

cm_ann = confusion_matrix(y_test, y_pred_ann)

plt.figure(figsize=(6, 6))

sns.heatmap(cm_ann, annot=True, fmt='d', cmap='Blues', cbar=False)

plt.title('ANN Confusion Matrix')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.show()

# ROC Curve for ANN

y_test_bin = label_binarize(y_test, classes=[0, 1, 2])

y_pred_bin = label_binarize(y_pred_ann, classes=[0, 1, 2])

fpr = dict()

tpr = dict()
```

```
roc_auc = dict()
for i in range(3):
    fpr[i], tpr[i], _ = roc_curve(y_test_bin[:, i], y_pred_bin[:, i])
    roc_auc[i] = auc(fpr[i], tpr[i])
plt.figure(figsize=(8, 6))
for i in range(3):
    plt.plot(fpr[i], tpr[i], label=f'Class {i} (AUC = {roc_auc[i]:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ANN ROC Curve')
plt.legend(loc='lower right')
plt.show()

print("="*50)

# Display final accuracy scores
accuracy_scores
```

Plagiarism report last page:



Shreya Proj paper

by Neelakantappa M

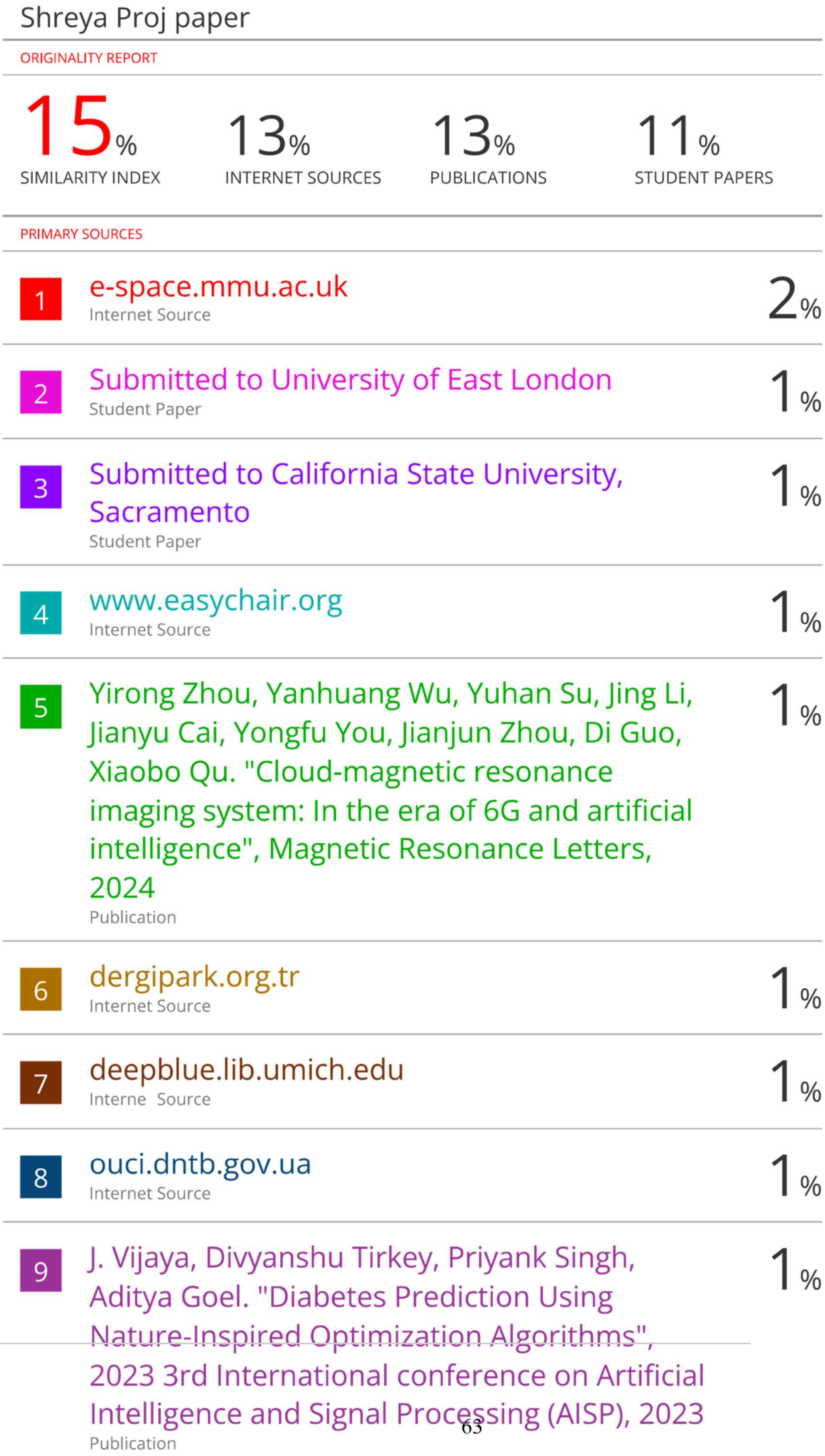
Submission date: 16-Apr-2025 03:27PM (UTC+0530)

Submission ID: 2647822681

File name: C_21_paper_draft_1_.pdf (874.91K)

Word count: 5222

Character count: 32626



Shreya Proj paper

ORIGINALITY REPORT

15%	13%	13%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	e-space.mmu.ac.uk Internet Source	2%
2	Submitted to University of East London Student Paper	1%
3	Submitted to California State University, Sacramento Student Paper	1%
4	www.easychair.org Internet Source	1%
5	Yirong Zhou, Yanhuang Wu, Yuhan Su, Jing Li, Jianyu Cai, Yongfu You, Jianjun Zhou, Di Guo, Xiaobo Qu. "Cloud-magnetic resonance imaging system: In the era of 6G and artificial intelligence", Magnetic Resonance Letters, 2024 Publication	1%
6	dergipark.org.tr Internet Source	1%
7	deepblue.lib.umich.edu Interne Source	1%
8	ouci.dntb.gov.ua Internet Source	1%
9	J. Vijaya, Divyanshu Tirkey, Priyank Singh, Aditya Goel. "Diabetes Prediction Using Nature-Inspired Optimization Algorithms", 2023 3rd International conference on Artificial Intelligence and Signal Processing (AISP), 2023 Publication	1%

Big Data Analytics In Diabetic Management System

1st Dr.M.Neelakantappa
Associate Professor
Department Of Information Technology
Vasavi College Of Engineering
Hyderabad, India
m.neelakanta@gmail.com

2nd Amaravadi Hasitha
Final Year B.E
Department Of Information Technology
Vasavi College of Engineering
Hyderabad, India
amaravdihasitha@gmail.com

3rd Chintapenta Shreya Sree
Final Year B.E
Department Of Information Technology
Vasavi College of Endineering
Hyderabad, India
cshreyasree12@gmail.com

***Abstract*—Diabetes is a long-term health condition that impacts millions of people worldwide,and correct and timely diagnosis is required to manage it effectively. In the present research study, we propose an improved model of diabetes prediction based on the usage of Big Data Analytics (BDA) and Machine Learning (ML) algorithms for enhancing the predictability of diabetes. Our research is based on the ideas developed in the reference paper, where the application of BDA in enabling healthcare decision-making was proposed. With the inclusion of the usage of advanced ML algorithms like Decision Trees, Random Forest, Logistic Regression,Naive Bayes, and Artificial Neural Networks (ANN), we enhance predictability for the diagnosis of diabetes.**

We employed a comprehensive healthcare dataset with various health metrics,including blood pressure, cholesterol, BMI, smoking, and physical activity. The data was heavily preprocessed, ranging from feature engineering to normalization, to get better model performance. All the models were trained and tested on accuracy, precision, recall, and ROC-AUC. Among the models, Random Forest and ANN were better with higher accuracy, detecting diabetic patients at risk effectively.

Our model not only increases prediction accuracy but also facilitates scalability with Big Data technology and practices. The ANN model, in particular, proved effective in finding complex patterns within the data, yielding a reliable solution towards real-time diabetes prediction. This effort illustrates the potential of integrating BDA and ML in healthcare for data-driven decision-making and supporting early diagnosis and management of diabetes. The model is deployed as an interactive user-friendly web application using Streamlit, where users can input health indicators and receive predictions from different models. Through its interactive interface, it facilitates healthcare professionals and patients by offering them real-time information, thereby ultimately improving diabetes management and patient care. Future developments focus on further model optimization, as well as on investigating additional features to further improve the prediction process.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Diabetes is a long-term disease that affects millions of people across the globe and leads to long-term health complications in the form of cardiovascular disease, kidney failure, and neuropathy if not treated. Since the incidence of diabetes is on the increase, its early diagnosis and best management are the need of the hour to prevent long-term health complications. Traditional detection methods, although efficient, fail to detect high-risk cases in the initial stage. Failure to achieve this necessitates the use of newer technologies in the form of Big Data Analytics (BDA) and Machine Learning (ML) to enhance prediction accuracy and trigger timely intervention.

The root article, The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making, highlights the revolutionary impact of BDA in healthcare by way of real-time analysis of data, predictive modeling, and the delivery of tailored treatments. With the processing of vast volumes of structured and unstructured healthcare data, BDA uncovers hidden patterns and relationships that result in better clinical decisions. Utilization of BDA in isolation is, however, not sufficient for making correct predictions, and thus the integration of ML models is required to augment the prediction process further.

II. Literature Survey

Big Data analytics improves healthcare with real-time tracking, predictive analysis, and AI-based decision-making. With challenges such as security and interoperability, new technologies such as blockchain and quantum computing hold the potential for greater efficiency and patient outcomes.The review summarizes various research paper that propose different apporaches for Big data analytics in health care.

Karatas et al. [1] explored Big Data's role in Healthcare Industry 4.0, integrating IoT, MCPS, and ML for optimized healthcare operations. The study highlighted benefits like predictive analytics and real-time monitoring while addressing challenges such as data privacy and cybersecurity. Future research suggests blockchain integration, AI-driven diagnostics, and improved security for seamless adoption.

Palanisamy et al. [2] examined Big Data analytics in healthcare, focusing on managing diverse data sources like electronic health records and medical imaging. The study highlighted machine learning, cloud computing, and security challenges while comparing healthcare frameworks. Future directions emphasized scalable, privacy-preserving solutions for improved decision-making and patient outcomes.

Nazir et al. [3] examined healthcare Big Data management, focusing on digital transformation through medical technology, electronic records, and wearable devices. The study addressed challenges like storage and cost efficiency while emphasizing predictive analytics and scientific programming for large-scale data processing. Future work aimed at optimizing decision-making frameworks for improved healthcare outcomes.

Wang et al. [4] explored Big Data analytics in logistics and supply chain management, categorizing supply chain analytics into descriptive, predictive, and prescriptive types. The study introduced a maturity framework, highlighting benefits like cost reduction and demand forecasting while addressing data complexity challenges. Future research focused on enhancing predictive models and real-time decision frameworks.

Wang et al. [5] explored Big Data analytics in healthcare, highlighting capabilities like pattern analysis, decision support, and predictive analytics. The study emphasized benefits such as improved decision-making and patient care while addressing challenges like data integration. It proposed strategies for effective implementation to enhance healthcare operations and organizational efficiency.

Bebortta et al. [6] proposed DeepMist, a mist computing framework using Deep Q Networks (DQN) for heart disease prediction. The study focused on reducing latency, improving accuracy, and optimizing energy efficiency. It outperformed benchmark models, addressing computational overhead. Future research aims to enhance mist computing resource allocation and secure healthcare data offloading.

Hasan et al. [7] developed a diabetes prediction framework using machine learning ensembles, combining AdaBoost and XGBoost with AUC-weighted soft voting. The study applied preprocessing techniques like outlier rejection and missing value imputation, achieving an AUC of 0.950. It highlighted the importance of data preprocessing for improving prediction accuracy and robustness.

Ahmed et al. [8] reviewed Big Data Analytics in healthcare, focusing on frameworks, applications, and challenges like data security and interoperability. The study highlighted AI-driven decision-making and predictive analytics for improved patient outcomes. Future research suggested blockchain for secure data management and real-time analytics for enhanced healthcare efficiency.

Hussain et al. [9] examined Big Data Analytics for improved clinical decision-making in healthcare. The study highlighted data processing techniques, security challenges, and real-time analytics. It emphasized the role of AI and machine learning in predictive healthcare, aiming to enhance patient care and decision-making frameworks.

Aceto et al. [10] examined ICTs in healthcare, highlighting e-health, m-health, and pervasive health paradigms. The study focused on advancements in data management, telemedicine, and health monitoring while addressing security and interoperability challenges. Future research emphasized AI, cloud computing, and IoT for improving healthcare efficiency and patient-centered services.

Dash et al. [11] explored Big Data in healthcare, focusing on data integration from hospital records, IoT, and biomedical research. The study highlighted challenges like security and interoperability while emphasizing AI-driven analytics, blockchain for secure data sharing, and quantum computing for efficient processing as future directions.

Wang et al. [12] examined biomedical Big Data analytics for precision health, focusing on multi-modal data integration, predictive modeling, and AI-driven decision support. The study addressed challenges like data quality and security while emphasizing translational bioinformatics and personalized healthcare as future directions.

Raghupathi et al. [13] examined Big Data analytics in healthcare, focusing on clinical decision support, disease surveillance, and population health management. The study discussed architectural frameworks, implementation methodologies, and challenges like data security and interoperability. Future directions emphasized AI-driven analytics, predictive modeling, and improved data-sharing frameworks for better healthcare outcomes.

Kankanhalli et al. [14] explored Big Data analytics in healthcare, emphasizing digitization, predictive analytics, and decision support. The study highlighted challenges like interoperability, privacy, and policy concerns while discussing real-time data processing. Future directions focused on AI-driven analytics and enhanced data-sharing frameworks for improved healthcare outcomes.

George et al. [15] surveyed Big Data analytics for diabetes prediction using machine learning algorithms like SVM, Random Forest, and Na¨ıve Bayes. The study highlighted data preprocessing, feature selection, and predictive analytics for improved accuracy. Future directions emphasized real-time analytics, optimized predictive models, and leveraging Big Data for early diabetes diagnosis.

Big Data analytics combines IoT, machine learning, and cloud computing to boost predictive analytics and decision-making in healthcare [1,2,5]. Researchers tackle issues such as security and interoperability [3,8,9], while the future directions are towards AI-driven diagnostics, blockchain, and quantum computing [6,11,12]. Mist computing and scalable solutions enhance healthcare efficiency [4,7,13,14,15].

III. Research Methodology

The research methodology used for the creation of the Diabetes Prediction and Management System Using Big Data Analytics and Machine Learning is a systematic four-stage approach. This methodology allows for systematic design, implementation,and verification of the suggested system, enhancing prediction accuracy and enabling real-time diagnosis through a web-based system.

A. Design Research

This phase entails an extensive review of diabetes prediction issues and the constraints of current models. The main activities are:

- Reviewing Related Literature: Discussing the base paper entitled ”The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making,” that emphasizes the importance of Big Data Analytics (BDA) in healthcare. The research finds loopholes in existing solutions, such as low accuracy, redundant features, and the requirement for effective predictive models.

- Identifying Gaps: Filling gaps in existing solutions, including restricted feature selection,overfitting issues, and inconsistent prediction outcomes. Current models frequently suffer from non-real time prediction capabilities, decreased precision, and the problem of having high-dimensional data.
- Defining System Objectives
  - Developing a diabetes prediction model with enhanced accuracy, utilizing a variety of AI/ML algorithms
  - Reducing model complexity by eliminating redundant features like 'Education','Income', and 'Veggies' from the dataset.
  - Providing a user-friendly **Streamlit web application** to enable real-time prediction.
  - Comparing model performance to select the best algorithm with optimized hyperparameters

B. *Conduct Research*

Following the definition of the research scope, technical details of the system were investigated and concluded:

- Dataset Selection: The Diabetes 012 Health Indicators BRFSS 2021 dataset was selected, consisting of 236,378 records with 22 original features. The dataset contains important indicators like 'HighBP', 'HighChol', 'BMI', 'Smoker', 'Stroke','HeartDiseaseorAttack', 'PhysActivity', and 'GenHlth'. To improve the model, unnecessary columns such as 'Education', 'Income', and 'Veggies' were removed to minimize noise and enhance predictive performance.
- Machine Learning Model Selection: Several ML algorithms were explored to predict the levels of diabetes risk efficiently:
  - Decision Tree (DT): To classify cases on the basis of feature split.
  - Random Forest (RF): To overcome overfitting by ensemble voting of multiple decision trees.
  - Logistic Regression (LR): For binomial and multi-class classification problems.
  - Naive Bayes (NB): A probabilistic model based on Bayes' Theorem.
  - Artificial Neural Network (ANN): To capture complex relationships and improve prediction performance.
- System Architecture Design: A robust architecture was developed to integrate various modules, including:
  - Data Preprocessing Module: For feature selection, normalization, and encoding.
  - Model Training and Evaluation Module: For training and testing various ML models.
  - Web Application Interface: A user-friendly interface developed using Streamlit for live predictions.

C. *Design Implementation*

The system was implemented iteratively with modularity and scalability in mind.This involved:

- Data Preprocessing and Model Training:
  - Data Preprocessing and Feature Engineering: Irrelevant features were dropped, and numeric data was normalized for better model performance.
  - Model Training: Various ML models were trained and validated to compare their performances with 80% of data for training and 20% for testing.
  - Hyperparameter Tuning: Grid Search and Random Search methods were employed to optimize models, specifically Random Forest and ANN models, which produced the best accuracy.
- Software Development and Model Deployment:
  - Streamlit Application: An interactive interface was developed using Streamlit, allowing real-time predictions from user inputs.
  - Model Integration: The models were all integrated into the application, with users being able to choose models and get dynamic predictions.

D. *Validation and Performance Evaluation*

Extensive evaluation and testing was done to verify the accuracy, performance, and usability of the system:

- Experimental Setup:
  - esting model performance using test datasets to determine robustness.
  - Validating model accuracy under different configurations and varying input conditions.
- Performance Metrics:
  - Accuracy: Assessed across multiple models to identify the best-performing model.
  - Precision and Recall: Ensured balanced detection of positive and negative instances.
  - F1-Score: Evaluated to balance precision and recall effectively.
  - ROC-AUC Score: Compared to evaluate the capacity of models to discriminate between classes.
- Model Comparison: The comparison indicated that Random Forest (RF) and Artificial Neural Network (ANN) models were superior to other models in terms of accuracy and generalization.
- User Feedback and Refinement:
  - nputs from the users were gathered to evaluate system usability and determine where improvements could be made.
  - Recursive enhancements were implemented to enhance the application interface and improve system response.

E. *Proposed System Architecture*

The system architecture includes:

- Data Preprocessing Layer: To perform cleaning, encoding, and normalization of data.
- Model Training and Validation Layer: For model training and testing performance.

- Web Application Layer: To have a user-friendly interface for users to engage with the prediction models.

Conclusion:This research approach guarantees a systematic, data-oriented, and user-centric method of building an efficient diabetes prediction system. The system, as proposed, was tested using strict testing with high accuracy and smooth real-time predictions. In the future, improvements will involve integrating real-time data using IoT integration, spreading to other diseases that are chronic in nature, and incorporating Explainable AI (XAI) to enhance model transparency and build user trust.

IV. Dataset Review

A. Dataset Overview:

- Dataset Name:Diabetes 012 Health Indicators BRFSS 2021
- Source:Behavioral Risk Factor Surveillance System (BRFSS), 2021
- File Name:‘diabetes_012_health_indicators_BRFSS2021.csv‘
- Number of Records (Rows):236,378
- Number of Features (Columns):22 (Before preprocessing)
- Dataset Size:Approximately 35 MB
- Data Format:CSV (Comma-Separated Values)
- Target Variable:‘Diabetes_012‘ (Multiclass classification)
  - ‘0‘ - No diabetes
  - ‘1‘ - Pre-diabetes
  - ‘2‘ - Diabetes

B. Data Collection and Purpose:

- he data was gathered through the **BRFSS 2021 Survey**, which compiles health-related information from the United States populace.
- It emphasizes different life habits, persistent health conditions, and healthcare availability.
- Objective: To enable research and analysis of diabetes-influencing factors and assist in predictive model development for early diagnosis and diabetes management.

C. Preprocessing and Feature Selection:

Features Retained After Preprocessing:Out of the initial 22 columns, 15 features were chosen after excluding irrelevant or less important features such as ’Diabetes_012’ ,’Education’ ,’Income’ ,’DiffWalk’ , ’NoDocbcCost’, ’Fruits’, and ’Veggies’.

The Table 1 describes about the dataset features

D. Target Variable Distribution

- Diabetes_012:
  - 0 (No Diabetes) – 70.2%
  - 1 (Pre-Diabetes) – 13.5%
  - 2 (Diabetes) – 16.3%

The data is slightly imbalanced with a greater number of non-diabetic patients than pre-diabetic and diabetic patients. To avoid issues due to imbalance, SMOTE (Synthetic Minority Over-sampling Technique) or other re-sampling methods can be employed during model training.

Feature Name	**Description**
HighBP	High Blood Pressure (0 = No, 1 = Yes)
HighChol	High Cholesterol (0 = No, 1 = Yes)
CholCheck	Cholesterol check in last 5 years (0 = No, 1 = Yes)
BMI	Body Mass Index (BMI) value
Smoker	Smoking status (0 = No, 1 = Yes)
Stroke	History of stroke (0 = No, 1 = Yes)
HeartDiseaseorAttack	History of heart disease or heart attack (0 = No, 1 = Yes)
PhysActivity	Physical activity in last 30 days (0 = No, 1 = Yes)
HvyAlcoholConsump	Heavy alcohol consumption (0 = No, 1 = Yes)
AnyHealthcare	Access to healthcare (0 = No, 1 = Yes)
GenHlth	General health rating (1 = Excellent to 5 = Poor)
MentHlth	Days with poor mental health in last 30 days
PhysHlth	Days with poor physical health in last 30 days
Sex	Gender (0 = Female, 1 = Male)
Age	Age group (1 = 18-24, 13 = 80+)

TABLE I
FEATURE DESCRIPTIONS

E. Preprocessing and Data Cleaning Steps

- Missing Values Handling: There are no missing values in the data.
- Feature Encoding: Label encoding was applied to transform categorical features into numerical values.
- Feature Scaling: Normalization was applied to continuous variables such as ‘BMI‘, ‘MentHlth‘, and ‘PhysHlth‘ to maintain the consistency of model performance.
- Feature Engineering: Further combinations of features (such as interaction between risk factors) were explored for enhanced predictive power.

F. Final Dataset Used to Train Models

- Rows: 236,378
- Columns After Preprocessing: 18 features + 1 target variable (‘Diabetes_012‘)

G. Rationale Behind Feature Deletion

- Education and Income: These columns were deleted in order to avoid bias during prediction and reduce model complexity. Even though socio-economic variables contribute towards healthcare accessibility, their deletion helped the models pay greater attention towards health indicators.
- Veggies: Vegetable consumption frequency was observed to bear little relation to diabetes incidence and was therefore omitted to avoid duplication.

Histogram plots give a graphical idea of how each feature is distributed in the dataset. Most binary categorical variables like HighBP, HighChol, CholCheck, Smoker, Stroke, HeartDiseaseorAttack, and PhysActivity exhibit class imbalance, where there is much higher frequency in one category (primarily 0), suggesting fewer people having those health conditions or dangerous behaviors. Variables like AnyHealthcare and

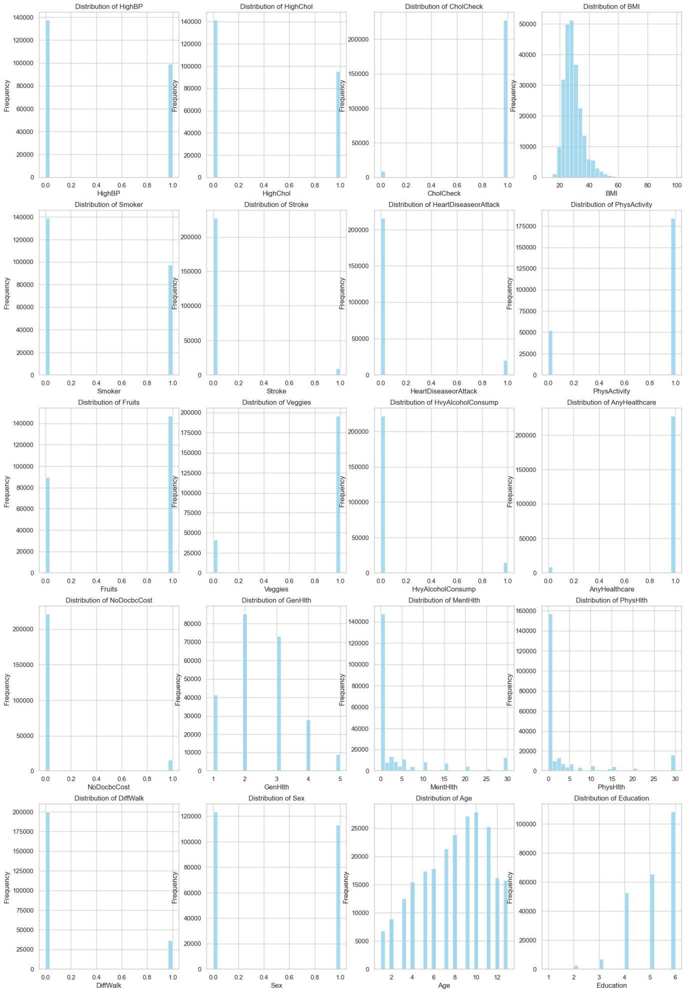


Fig. 1. Histogram Of Features

CholCheck indicate that most people have access to healthcare and have had cholesterol checks.

Continuous.variables like BMI are right-skewed with most values ranged from 20 to 40, which is characteristic of general populations. MentHlth and PhysHlth are also right-skewed with most individuals reporting fewer unhealthy days in both per month. Ordinal.variables like GenHlth have most responses in the middle range, indicating that most individuals perceive average health. Categorical attributes such as Age, Education, and Income are discretized into intervals, with relatively uniform distributions in some (e.g., Age) and non-uniform distributions in others (e.g., Education, biased toward higher values).

In general, the dataset shows both skewness and imbalance in some features, which are crucial factors to consider during data preprocessing and model training. These observations inform feature engineering, normalization, and class imbalance handling in predictive modeling.

The violin plots show the distribution of important features by diabetes status (0 = No, 1 = Prediabetes, 2 = Diabetes). Features like BMI, GenHlth, and PhysHlth have obvious variation between diabetic and non-diabetic groups. Increased BMI and worse general/physical health are more prevalent in diabetic groups. Income and Education are lower in diabetic groups, suggesting socioeconomic effect. MentHlth and SleepTime have moderate spread, indicating some effect but less stability. These plots indicate how lifestyle, health, and financial considerations relate to diabetes risk, giving good insight into feature relevance.

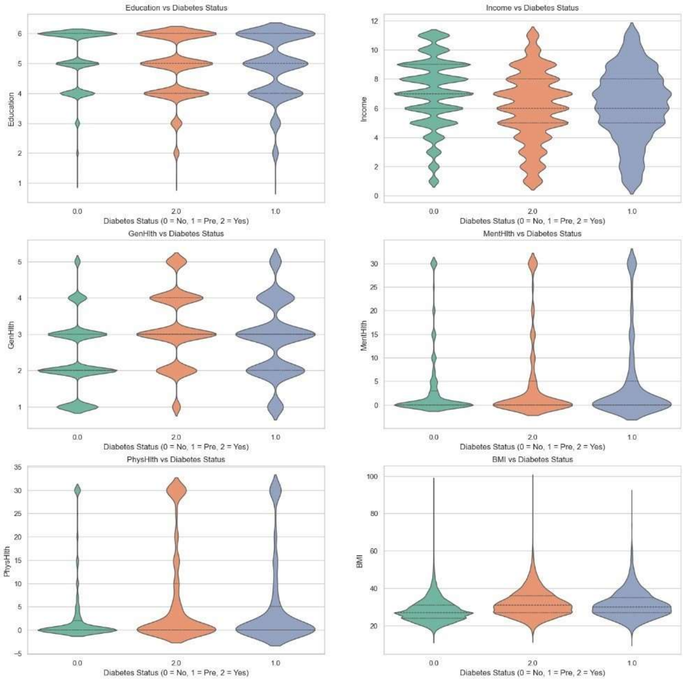


Fig. 2. Comparision of Diabetes Status with different Features using Vilon Plots

V. Proposed Work

The proposed work will be to design an effective and precise Diabetes Prediction System based on machine learning and deep learning algorithms. The app will employ multiple algorithms, such as Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and an Artificial Neural Network (ANN), to offer a thorough diagnosis based on health indicators input by the user. The ultimate goal is to design a simple, accessible tool for early diagnosis and treatment.

A. Objectives

- Create an interactive web-based interface with Streamlit.
- Use several prediction models to allow for comparative analysis.
- Increase the accuracy of predictions by employing ensemble learning strategies.
- Provide robustness and scalability to the application.
- Offer interpretable results to assist in making well-informed medical decisions.

B. Methodology

- Data collection and preprocessing:
  - Use a public diabetes dataset that includes health indicators.
  - Conduct data cleaning, normalization, and categorical variable encoding.
  - Manage missing values and validate data consistency.
- Model Selection and Training:
  - Train Decision Tree, Random Forest, Logistic Regression, Naive Bayes, and ANN models with optimal hyperparameter tuning.
  - Use accuracy, precision, recall, F1-score, and ROC-AUC as model evaluation metrics.
- Model Deployment:
  - Deploy the trained models using Streamlit.
  - Provide users with an intuitive interface to input health parameters.
  - Perform real-time predictions and display results using the best-performing model.
- Performance Comparison:
  - Enable users to compare predictions across different models.
  - Provide graphical visualizations for better understanding.

C. Expected Outcome

- Precise early diagnosis of diabetes through state-of-the-art machine learning and deep learning methods.
- Improved decision-making for healthcare providers.
- User-friendly platform for users to track their health condition effectively.

This suggested work will assist in proactive management of healthcare by detecting vulnerable individuals and suggesting early interventions, thus decreasing the load of diabetes complications.

VI. System Architecture

A. Data Preprocessing Layer

This layer is tasked with prepping the dataset for training and prediction using the model. Major activities include:

- *Data Cleaning:Redundant and irrelevant features such as ‘Education‘, ‘Income‘, ‘Veggies‘,’Diffwalk’, ’Fruits’, and ’NoDocbcCost’ were dropped to remove noise and enhance model efficiency.
- Feature Selection and Encoding:Categorical attributes like ‘Sex‘, ‘HighBP‘, ‘HighChol‘, and ‘Smoker‘ were encoded to numerical values with label encoding.
- Data Normalization:Continuous attributes like ‘BMI‘, ‘MentHlth‘, and ‘PhysHlth‘ were normalized to have uniformity in all the features.

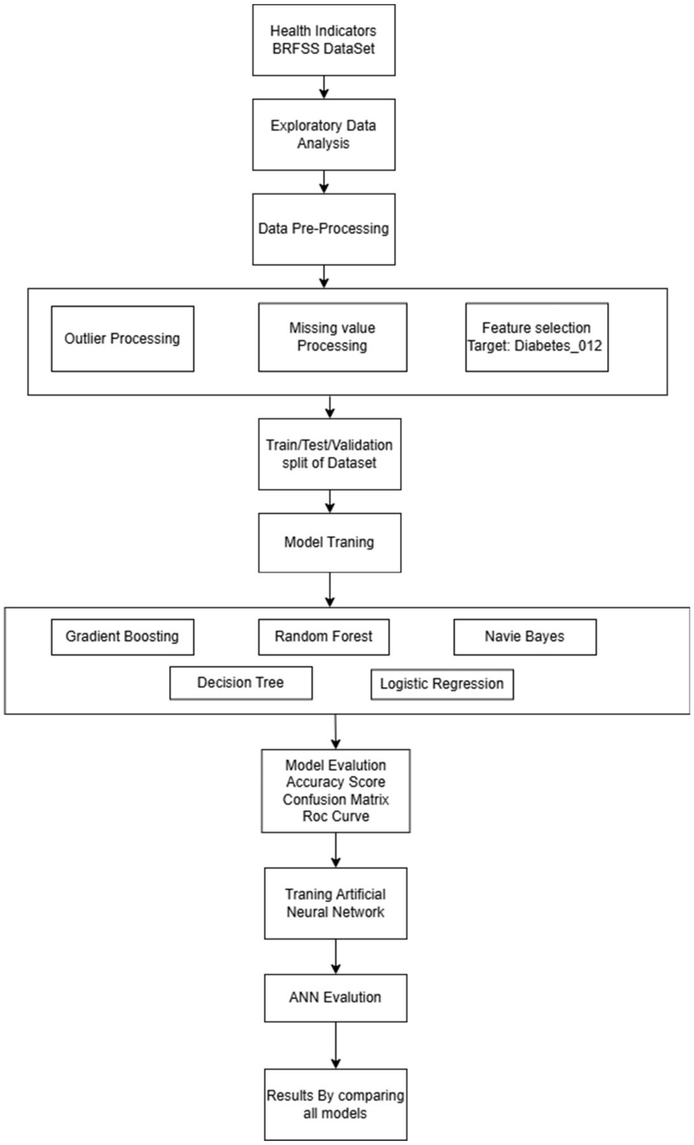

Health Indicators BRFSS DataSet
Exploratory Data Analysis
Data Pre-Processing
Outlier Processing
Missing value Processing
Feature selection Target: Diabetes_012
Train/Test/Validation split of Dataset
Model Traning
Gradient Boosting
Random Forest
Navie Bayes
Decision Tree
Logistic Regression
Model Evalution Accuracy Score Confusion Matrix Roc Curve
Traning Artificial Neural Network
ANN Evalution
Results By comparing all models

Fig. 3. Architecture Diagram

B. Model Training and Validation Layer

This layer is tasked with training, fine-tuning, and validating various machine learning models for high prediction accuracy. Some of the major processes are:

- Model Implementation: The following models were trained and tested:
  - Decision Tree (DT)
  - Random Forest (RF)
  - Logistic Regression (LR)
  - Naive Bayes (NB)
  - Artificial Neural Network (ANN)
- Hyperparameter Tuning:Grid Search and Random Search methods were used to tune model parameters for Random Forest and ANN, enhancing prediction accuracy.
- Model Evaluation: Performance of every model was checked using the following metrics:

- Accuracy
- Precision and Recall
- F1-Score
- ROC-AUC Score

C. Web Application Layer

The web application, developed with Streamlit, is an interactive interface through which users can provide health-related input parameters and obtain predictions in real time. The main features are:

- User Input Interface:The user inputs values for important health parameters like ‘BMI‘, ‘HighBP‘,‘HighChol‘, and other relevant factors via a sidebar interface.
- Model Selection and Prediction:Users select among various models, and the application performs the input data processing to return a prediction regarding diabetes status (‘No Diabetes‘, ‘Pre-Diabetes‘, ‘Diabetes‘).
- Result Display:The system returns prediction results with high confidence and accuracy scores to ensure credible and real-time decision-making.

D. Cloud Database and Model Storage

- Data Storage:Preprocessed data and trained models are stored in a cloud database in a readily retrievable manner for inference.
- Model Deployment:Trained models are dynamically loaded for real-time prediction and system scalability.

E. Data Flow and System Workflow

- User inputs are received and forwarded to the preprocessing layer.
- Preprocessed data is passed to the chosen model for prediction.
- Predicted outputs are shown through the web application in real time.

This architecture provides a smooth, precise, and efficient diabetes prediction system that can efficiently handle large-scale health data.

VII. Methodology

A. Introduction

The research in this paper is intended to predict the probability of diabetes through machine learning models. The methodology includes data collection, preprocessing, model selection, training, evaluation, and deployment through a user-friendly interface developed with Streamlit.

B. Data Collection

The data used in this research is derived from publicly released health indicator datasets such as the Behavioral Risk Factor Surveillance System (BRFSS). It comprises different health measurements such as blood pressure, cholesterol, BMI, smoking status, and physical activity.

C. Data Preprocessing

- Data Cleaning: Dealing with missing values and deleting duplicates.
- Feature Engineering: Generating useful features from available data.
- Normalization: Feature scaling for uniform model input.
- Data Splitting: Splitting data into training and test sets with an 80-20 split.

D. Model Selection

The research compares five models for diabetes prediction:

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression
- Naive Bayes Classifier
- Artificial Neural Network (ANN)

E. Model Training and Evaluation

- Models are trained on the training dataset.
- Metrics used for evaluation are Accuracy, Precision, Recall, and F1-Score.
- The ANN model is optimized with suitable hyperparameters using TensorFlow.

F. Prediction Mechanism

The health parameters are input by the user via the Streamlit interface. The input is forwarded to the chosen model, and predictions are generated. For the ANN model, predictions are accessed via the np.argmax function.

G. Deployment

- The models are Pickled for fast loading and inference.
- ANN is saved in .h5 format via tensorflow.keras.models.save_model.
- A Streamlit web application has a simple and friendly interface for prediction.

The research gives a comparative evaluation of several machine learning models for prediction of diabetes with an easily implementable solution for healthcare professionals to take informed data-driven decisions. The ANN model should exhibit improved accuracy and robustness as it can learn more complex things. Future enhancements can include employing more sophisticated models such as XGBoost, ensemble learning, and the incorporation of real-time data from wearable health devices.

VIII. Results And Conclusion

The bar chart below gives a comparative assessment of different classification models by their accuracy scores when used to predict diabetes. Among the models, Gradient Boosting, Artificial Neural Networks (ANN), and Logistic Regression were all able to attain the maximum accuracy of 0.84, which testifies to their high performance and applicability for this classification problem. These models were able to capture the intricate patterns in the data and generalize well on new data.

S.No	ML Model	Accuracy Score
1	Decision Tree	75%
2	Random Forest	83%
3	Gradient Boosting	84%
4	Logistic Regression	83%
5	Naive Bayes	75%
6	ANN (Artificial Neural Network)	84%

TABLE II

ACCURACY COMPARISON OF DIFFERENT CLASSIFICATION MODELS

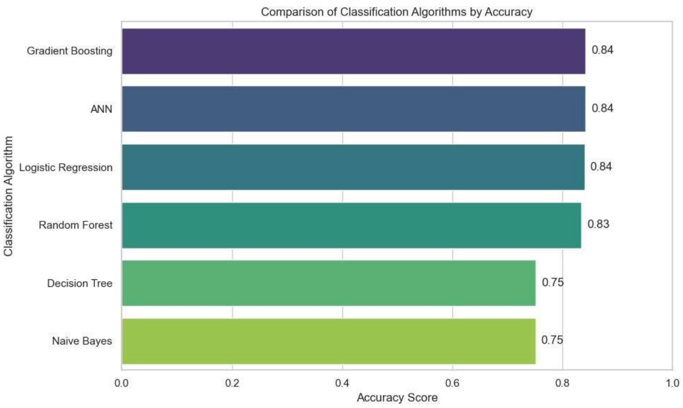


Fig. 4. Comparison of accuracy with different classification algorithms'.

Random Forest also did well with an accuracy of 0.83, and this makes it a strong and stable model because it uses ensemble learning. Decision Tree and Na¨ıve Bayes, although less complex and quicker to train, were behind with an accuracy of 0.75. This indicates that they might not be as good at detecting the subtle relationships in the dataset as more complicated models.

This contrast emphasizes the significance of model selection in obtaining the best predictive performance. It also indicates that ensemble and deep learning techniques offer a considerable benefit in processing structured health data for classification problems. In deployment or further optimization, concentrating on the highest-performing models (Gradient Boosting, ANN, and Logistic Regression) would be the most efficient strategy.

The heatmap above displays the correlation matrix between all the numeric features of the dataset except the target variable. The correlation values vary between -1 (strong negative) and +1 (strong positive), with the values close to 0 indicating little or no linear relationship.

The diagonal elements are all 1.0 since each attribute is maximally correlated with itself. Of the attributes, GenHlth is highly positively correlated with PhysHlth (0.42) and DiffWalk (0.41), indicating that those with poorer general health also report more physically unhealthy days and walking difficulties. Likewise, HighBP and HighChol are moderately correlated (0.28), consistent with known medical connections.

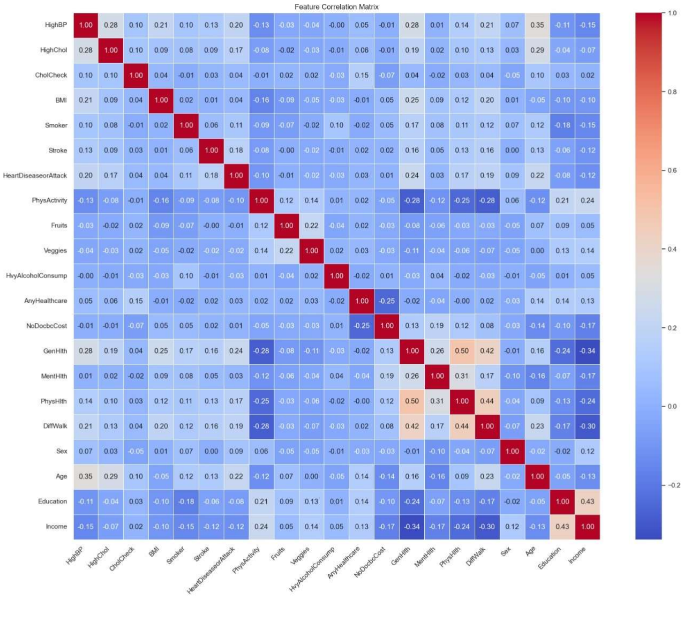


Fig. 5. Correlation matrix

Negative correlations between PhysActivity and DiffWalk (–0.28), MentHlth and GenHlth (–0.34), and Income and NoDocbcCost (–0.24) were seen, showing that increased income lowers cost-related access barriers to healthcare. Positive correlations between Age and HighBP and between Age and HighChol were seen, as would be expected given age-associated health threats.

Most feature pairs, however, demonstrate weak correlations (–0.1 values 0.1), indicating minimal multicollinearity. This guarantees that the majority of features make distinct contributions to the model and lends support for the application of multivariate methods in prediction.

IX. CONCLUSION AND FEATURE SCOPE

A. Conclusion

The research gives a comparative evaluation of several machine learning models for prediction of diabetes with an easily implementable solution for healthcare professionals to take informed data-driven decisions. The ANN model should exhibit improved accuracy and robustness as it can learn more complex things.

B. Future Work

Future enhancements can include employing more sophisticated models such as XGBoost, ensemble learning, and the incorporation of real-time data from wearable health devices. This correlation analysis is crucial to feature selection and identifying variable interactions within the dataset.

REFERENCES

[1] Karatas, Mumtaz, et al. ”Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives.” Expert Systems with Applications 200 (2022): 116912.

[2] Palanisamy, Venketesh, and Ramkumar Thirunavukarasu. ”Implications of big data analytics in developing healthcare frameworks–A review.” Journal of King Saud University-Computer and Information Sciences 31.4 (2019): 415-425.

[3] Nazir, Shah, et al. ”A comprehensive analysis of healthcare big data management, analytics and scientific programming.” IEEE Access 8 (2020): 95714-95733.

[4] Wang, Gang, et al. ”Big data analytics in logistics and supply chain management: Certain investigations for research and applications.” International journal of production economics 176 (2016): 98-110.

[5] Wang, Yichuan, LeeAnn Kung, and Terry Anthony Byrd. ”Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations.” Technological forecasting and social change 126 (2018): 3-13.

[6] Bebortta, Sujit, et al. ”DeepMist: Toward deep learning assisted mist computing framework for managing healthcare big data.” IEEE Access 11 (2023): 42485-42496.

[7] Hasan, Md Kamrul, et al. ”Diabetes prediction using ensembling of different machine learning classifiers.” IEEE Access 8 (2020): 76516-76531.

[8] Ahmed, Awais, et al. ”Harnessing big data analytics for healthcare: A comprehensive review of frameworks, implications, applications, and impacts.” IEEE Access 11 (2023): 112891-112928.

[9] Hussain, Fatima, et al. ”Leveraging big data analytics for enhanced clinical decision-making in healthcare.” IEEE Access 11 (2023): 127817-127836.

[10] Aceto, Giuseppe, Valerio Persico, and Antonio Pescape´. ”The role of Information and Communication Technologies in healthcare: taxonomies, perspectives, and challenges.” Journal of Network and Computer Applications 107 (2018): 125-154.

[11] Dash, Sabyasachi, et al. ”Big data in healthcare: management, analysis and future prospects.” Journal of big data 6.1 (2019): 1-25.

[12] Dash, Sabyasachi, et al. ”Big data in healthcare: management, analysis and future prospects.” Journal of big data 6.1 (2019): 1-25.

[13] Raghupathi, Wullianallur, and Viju Raghupathi. ”Big data analytics in healthcare: promise and potential.” Health information science and systems 2 (2014): 1-10.

[14] Kankanhalli, Atreyi, et al. ”Big data and analytics in healthcare: Introduction to the special section.” Information Systems Frontiers 18 (2016): 233-235.

[15] George, AS Hovan, et al. ”A Survey study on big data analytics to predict diabetes diseases using supervised classification methods.” Partners Universal International Innovation Journal 1.1 (2023): 1-8.

[16] Eswari, T., P. Sampath, and S. J. P. C. S. Lavanya. ”Predictive methodology for diabetic data analysis in big data.” Procedia Computer Science 50 (2015): 203-208.

[17] Kolesnichenko, Olga, et al. ”Big data analytics of inpatients flow with diabetes mellitus type 1: Revealing new awareness with advanced visualization of medical information system data.” 2019 9th International Conference on Cloud Computing, Data Science Engineering (Confluence). IEEE, 2019.

[18] Collins, Gary S., et al. ”Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting.” BMC medicine 9 (2011): 1-14.

[19] Bhotta, Dinakar, et al. ”An investigation into usability of big data analytics in the management of Type 2 Diabetes Mellitus.” 24th Annual Conference of the Asia Pacific Decision Sciences Institute: Full papers. University of Southern Queensland, 2019.

[20] Khanra, Sayantan, et al. ”Big data analytics in healthcare: a systematic literature review.” Enterprise Information Systems 14.7 (2020): 878-912.

[21] Macinati, Manuela S., and Eugenio Anessi-Pessina. ”Management accounting use and financial performance in public health-care organisations: Evidence from the Italian National Health Service.” Health Policy 117.1 (2014): 98-111.

[22] Kamiran, Faisal, and Toon Calders. ”Data preprocessing techniques for classification without discrimination.” Knowledge and information systems 33.1 (2012): 1-33.

[23] Perveen, Sajida, et al. ”Performance analysis of data mining classification techniques to predict diabetes.” Procedia Computer Science 82 (2016): 115-121.

[24] Shyni, S., R. Shantha Mary Joshitta, and L. Arockiam. ”Applications of big data analytics for diagnosing diabetic mellitus: issues and challenges.” International Journal of Recent Trends in Engineering Research 2.06 (2016): 454-461.

[25] Nagarajan, Srideivanai, and R. M. Chandrasekaran. ”Design and implementation of expert clinical system for diagnosing diabetes using data mining techniques.” Indian Journal of science and Technology 8.8 (2015): 771-6.