**A MINOR PROJECT REPORT**

ON

**Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19**

*Submitted in partial fulfilment of the requirement for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**IN**

## COMPUTER SCIENCE AND ENGINEERING

**(DATA SCIENCE)**

BY

### Katravath Venkatesh 21P61A6794

### Lakavath Yaku 21P61A67A9

### Kamuni Sanjay Kumar 21P61A6784

*Under the esteemed guidance of*

***Mr. Krishna Chaitanya***

***Assistant Professor***

***Dept. of CSD***



## VIGNANA BHARATHI INSTITUTE OF TECHNOLOGY

(A UGC Autonomous Institution, Approved by AICTE, Affiliated to JNTUH, Accredited by

NBA & NAAC) Aushapur (V), Ghatkesar (M), Medchal(dist)

**December – 2024**



Aushapur (V), Ghatkesar (M), Hyderabad, Medchal – Dist, Telangana – 501 301.

## DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)

***CERTIFICATE***

*This is to certify that the mini project titled “***Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19*”*** *submitted by* ***Katravath Venkatesh(21P61A6794), Lakavath Yaku(21P61A67A9), Kamuni Sanjay Kumar(21P61A6784)*** *in B.Tech IV-II semester Computer Science & Engineering(Data Science) is a record of the bonafide work carried out by them. The results embodied in this report have not been submitted to any other University for the award of any degree****.***

|  |  |  |
| --- | --- | --- |
| **INTERNAL GUIDE** |  | **PROJECT CO-ORDINATOR** |
| ***Mr. Krishna Chaitanya***  **(Assistant Professor)** |  | **Dr. P.Punitha (Assistant Professor)** |
| **HEAD OF THE DEPARTMENT** |  | **EXTERNAL EXAMINER** |
| **Dr. Y. Raju** |  |  |

**(Associate Professor)**

### DECLARATION

We, **Katravath Venkatesh, Lakavath Yaku, Kamuni Sanjay Kumar,** bearing hall ticket numbers **21P61A6794, 21P61A67A9, 21P61A6784** hear by declare that the mini project report entitled “**Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19**” under the guidance of **Mr. Krishna Chaitanya**, Department of Computer Science & Engineering (Data Science), **Vignana Bharathi Institute of Technology, Hyderabad,** have submitted to Jawaharlal Nehru Technological University Hyderabad, Kukatpally, in partial fulfilment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering(Data science).

This is a record of bonafide work carried out by us and the results embodied in this project have not been reproduced or copied from any source. 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.

**Katravath Venkatesh (21P61A6794)**

**Lakavath Yaku (21P61A67A9) Kamuni Sanjay Kumar (21P61A6784)**

**ACKNOWLEDGEMENT**

We are extremely thankful to our beloved Chairman, **Dr. N. Goutham Rao** and secretary, **Dr. G. Manohar Reddy** who took keen interest to provide us the infrastructural facilities for carrying out the project work. Self-confidence, hard work, commitment and planning are essential to carry out any task. Possessing these qualities is sheer waste, if an opportunity does not exist. So, we whole- heartedly thank **Dr. P. V. S. Srinivas**, Principal, and **Dr. Y. Raju**, Head of the Department, Computer Science and Engineering (Data science) for their encouragement, support and guidance in carrying out the project.

We would like to express our indebtedness to the project coordinator, **Mr. Krishna Chaitanya**, Assistant Professor, Department of CSE (Data science) for her valuable guidance during the course of project work.

We thank our Project Guide, **Mr. Krishna Chaitanya**, Assistant Professor, for providing us with an excellent project and guiding us in completing our major project successfully.

We would like to express our sincere thanks to all the staff of Computer Science and Engineering (Data science), VBIT, for their kind cooperation and timely help during the course of our project. Finally, we would like to thank our parents and friends who have always stood by us whenever we were in need of them.

### ABSTRACT

The COVID-19 pandemic has significantly impacted global health, leading to an increased incidence of cardiovascular complications. This study aims to develop a comparative heart stroke prediction system using machine learning techniques to analyze and predict stroke risk, examining trends before and after the onset of COVID-19. Leveraging data-driven insights, the system explores potential shifts in risk factors, including demographics, comorbidities, lifestyle changes, and healthcare access, that have evolved due to pandemic-related stress and health disruptions. The proposed model utilizes supervised machine learning algorithms, including logistic regression, decision trees, and neural networks, trained on pre-COVID and post-COVID datasets. The analysis seeks to reveal any notable variations in stroke prediction accuracy, sensitivity, and specificity across both datasets. Feature importance analysis highlights key predictors, offering insights into new risk factors that emerged post-COVID. The results could potentially inform healthcare providers about evolving cardiovascular risks and assist in prioritizing high-risk patients for early interventions. By providing an integrated platform to track changes in stroke risk predictors, this study underscores the importance of adapting predictive health models to account for pandemic-related effects on public health.

**Keywords:** Heart stroke prediction, Machine learning, COVID-19 impact, Pre- and post-pandemic analysis, Healthcare access, Cardiovascular risk factors.

**DEPARTMENT OF**

**COMPUTER SCIENCE AND ENGINEERING**

## (Data Science)

### VISION

To be recognized as a Centre of Excellence in Data Science to meet the ever-growing needs of Industry and Society.

### MISSION

* To empower students with innovative and cognitive skills to gain expertise in the field of Data science.
* To Inculcate the seeds of knowledge by providing industry conducive environment to enable students excel in the field of Data Science.
* To provide an appropriate ambience to nurture the young Data Science professionals.

#### PROGRAM EDUCATIONAL OBJECTIVES (PEOs)

**PEO 1: Domain Knowledge**: Develop a broad academic and practical literacy in computer science, statistics, and optimization, with relevance in data science.

**PEO 2: Professional Employment**: Employed in industry government and entrepreneurial endeavors to have a successful professional career.

**PEO 3: Higher Degrees**: Pursue higher education in the domain of data analytics or research.

**PEO 4: Engineering Citizenship**: Contribute to the society and human well-being by applying ethical principles.

**PEO 5: Lifelong Learning:** Pursue lifelong learning in generating innovation engineering research-based solution using latest innovation tolls and technologies.

### PROGRAM OUTCOMES (POs)

**Engineering graduates will be able to:**

1. **Engineering Knowledge**: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
2. **Problem Analysis**: Identify, formulate, review research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.
3. **Design/development of solutions**: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, and environmental considerations.
4. **Conduct investigations of complex problems**: Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
5. **Modern tool usage**: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.
6. **The engineer and society**: Apply reasoning informed by contextual knowledge to assess societal, health, safety, legal and cultural issues, and the consequent responsibilities relevant to professional engineering practice.
7. **Environment and sustainability**: Understand the impact of professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.
8. **Ethics**: Apply ethical principles and commit to professional ethics and responsibilities and norms of engineering practice.
9. **Individual and teamwork:** Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.
10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend

and write effective reports and design documentation, make effective Presentations, and give and receive clear instructions.

1. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one’s own work, as a member and leader in a team, to manage projects and in multidisciplinary Environments.
2. **Life-long learning**: Recognize the need for and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

### PROGRAM SPECIFIC OUTCOMES (PSOs)

**PSO1:** Understand fundamental concepts in statistics, mathematics and computer Science to gain an understanding and working knowledge of various tools for analysis.

**PSO2:** Represent the knowledge, predicate logic and then transform the real-life information into visually appealing data using suitable tools.

**PSO3:** Get Expertise in different aspects and appropriate models of Data Science and use large data sets to cater to the growing demand for data scientists and engineers in industry.

#### Course Outcomes (COs)

**CO1** - Identify the problem by applying acquired knowledge from survey of technical publications

**CO2** - Analyze and categorize identified problem to formulate and fine the best solution after considering risks.

**CO3** - Choose efficient tools for designing project.

**CO4** - Build the project through effective team work by using recent technologies.

**CO5** - Elaborate and test the completed task and compile the project report.

#### Correlation Levels

|  |  |
| --- | --- |
| Substantial/ High | 3 |
| Moderate/ Medium | 2 |

#### CO – PSO Correlation Matrix

|  |  |  |  |
| --- | --- | --- | --- |
| **COs** |  | **PSOs** |  |
| **PSO1** | **PSO2** | **PSO3** |
| **CO1** | 2 | 3 | 2 |
| **CO2** | 2 | 3 | 2 |
| **CO3** | 2 | 2 | 3 |
| **CO4** | 3 | 2 | 3 |
| **CO5** | 1 | 2 | 2 |
| **CO** | 2 | 2.4 | 2.4 |

#### CO – PO Correlation Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **COs** | **POs** | | | | | | | | | | | |
| **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** |
| **CO1** | 1 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 1 | 2 | 3 |
| **CO2** | 2 | 3 | 2 | 2 | 2 | 3 | 1 | 1 | 2 | 3 | 2 | 2 |
| **CO3** | 3 | 2 | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 2 | 3 | 2 |
| **CO4** | 2 | 2 | 3 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 2 | 2 |
| **CO5** | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 |
| **CO** | 2.2 | 2.4 | 2.6 | 1.8 | 2 | 2.4 | 2 | 1.6 | 1.8 | 2 | 2.2 | 2.2 |

#### Project Outcomes (PROs)

1. **Real-time Interactive Art Installation:** Create an interactive art installation where users' facial movements and gestures are tracked in real-time using OpenCV. Users can see their movements translated into vibrant brush strokes on a virtual canvas, allowing them to create dynamic paintings with their gestures and expressions.
2. **Educational Tool for Gesture Recognition:** Develop an educational tool that uses facial and object movement detection to teach students about computer vision concepts. Users can see how their facial expressions and hand movements are interpreted by the system and translated into visual representations on a virtual canvas, helping them understand the underlying principles of gesture recognition algorithms.
3. **Therapeutic Painting Application:** Design a therapeutic painting application that utilizes facial movement detection to assist individuals with motor disabilities in expressing themselves through art. By tracking subtle facial gestures and translating them into brush strokes on a virtual canvas, users can engage in creative expression and improve their motor skills in a supportive digital environment.
4. **Virtual Collaborative Art Platform:** Build a virtual collaborative art platform where multiple users can paint together on a shared canvas using their facial movements and object interactions. OpenCV is used to detect and track users' faces and gestures, allowing them to contribute to the artwork in real-time from different locations. This platform fosters creativity and collaboration among users, regardless of physical distance.
5. **Interactive Advertising Display:** Create an interactive advertising display that captures the attention of passers by responding to their facial expressions and movements. OpenCV is employed to detect faces and analyze expressions, triggering visual effects and animations on a virtual canvas that correspond to users' actions. This innovative advertising solution engages audiences in a personalized and immersive experience, increasing brand engagement and customer interaction.

**PRO – PSO Correlation Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
| **PROs** |  | **PSOs** |  |
| **PSO1** | **PSO2** | **PSO3** |
| **PRO1** | 2 | 2 | 3 |
| **PRO2** | 1 | 2 | 2 |
| **PRO3** | 1 | 1 | 1 |
| **PRO4** | 2 | 2 | 2 |
| **PRO5** | 1 | 1 | 2 |
| **PRO** | 1.4 | 1.6 | 2 |

#### PRO – PO Correlation Matrix

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **PROs** |  |  |  |  | **POs** | | |  |  |  |  |  |
| **PO1** | **PO2** | **PO3** | **PO4** | **PO5** | **PO6** | **PO7** | **PO8** | **PO9** | **PO10** | **PO11** | **PO12** |
| **PRO1** | 1 | 2 | 2 | 2 | 3 | 2 | 2 | 1 | 2 | 3 | 2 | 2 |
| **PRO2** | 2 | 2 | 3 | 3 | 3 | 2 | 3 | 2 | 2 | 2 | 2 | 3 |
| **PRO3** | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 2 | 2 | 3 | 2 | 3 |
| **PRO4** | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 1 | 2 | 3 | 2 | 2 |
| **PRO5** | 3 | 1 | 2 | 2 | 3 | 3 | 1 | 2 | 2 | 2 | 3 | 2 |
| **PRO** | 2 | 1.8 | 2.4 | 2.2 | 2.6 | 2.6 | 2.2 | 1.6 | 2 | 2.6 | 2.2 | 2.4 |

**TABLE OF CONTENTS**

**CHAPTER PAGE NO**

**Abstract**

**List of Figures**

1. **Introduction**  2
   1. Existing System 3
   2. Proposed System 4
   3. Aim and Objective 5
   4. Scope 6
2. **Literature Survey** 8
3. **Design**  14
   1. Software Requirements 14
   2. Functional Requirements 15
   3. Non-Functional Requirements 18
   4. Model Architecture 21
4. **Implementation**  23
5. **Results**  33
6. **Testing** 43
   1. System Testing 43
   2. Software Testing Strategies 43
   3. Testing 44

6.3.1 Testing Methods 44

* + - 1. Define the baseline models (Pre-COVID) 44
      2. Introduction Post COVID Risk Factors 44
      3. Model Adapton and Training 44
      4. Performance Evaluation and Comparison 45
      5. Interpretation of Results and Insights 45

1. **Conclusion**  47
2. **Future Enhancement** 49
3. **References**  51
4. **Conference Paper** 55

|  |  |  |
| --- | --- | --- |
| **Fig No** | **Figure Name** | **Page No** |
| 3.4 | System Architecture | 21 |
| 5.1 | Smoking Status | 34 |
| 5.2 | Glucose levels | 34 |

**List of Figures**

**CHAPTER-1**

## 1. Introduction

The COVID-19 pandemic has significantly impacted healthcare, reshaping our understanding of cardiovascular conditions such as heart disease and stroke. With the increasing importance of health data analytics, machine learning has proven highly effective in predicting cardiovascular disease risks by analyzing datasets that encompass a range of risk factors, including age, blood pressure, cholesterol levels, lifestyle choices, and pre-existing health conditions. Given that heart disease and stroke are major global causes of mortality, studying these conditions through a comparative machine learning approach—focusing on insights from pre- and post-COVID-19 periods could help identify shifts in disease patterns, risk factors, and outcomes due to pandemic-related changes like prolonged inactivity, heightened stress, and the virus's direct impact on the cardiovascular system.

The objective of this analysis is to determine if and how traditional predictors of heart disease and stroke risk have changed post-COVID-19. Specifically, it aims to uncover new risk factors associated with COVID-19, analyze any shifts in the impact or prevalence of established risk factors (e.g., hypertension, diabetes), and assess whether existing models need retraining or entirely new models to accommodate these changes. Data sources include health records, biometric data, and lifestyle information from both before and after the pandemic. Various machine learning algorithms, such as logistic regression, decision trees, random forests, and neural networks, are applied to develop and compare prediction models across these periods, helping to reveal evolving patterns in risk factors and disease trends.

This analysis presents several challenges, including the availability of high-quality post-pandemic health data and the difficulty of isolating the effects of COVID-19 from other lifestyle changes. Additionally, the models must be designed carefully to ensure they do not introduce biases, especially given the unequal impact of the pandemic across different demographic groups. Insights from this comparative study could be transformative, helping to inform public health strategies, guide personalized treatment approaches, and enable healthcare providers to address new vulnerabilities within post-pandemic populations.

* 1. **Existing System:**

The impact of COVID-19 on heart stroke prediction has introduced significant changes to how we assess risks and develop predictive models in healthcare. Before the pandemic, models focused on traditional risk factors such as age, hypertension, diabetes, and lifestyle, relying on standardized medical records that provided a stable basis for predictive analytics. However, COVID-19 has brought new considerations, such as vascular complications, heightened inflammation, and potential clotting issues, particularly in individuals recovering from the virus. Additionally, the effects of "long COVID," which has been associated with heart and vascular health issues, have reshaped our understanding of stroke risk. This shift has necessitated updates to machine learning models, which previously depended on traditional factors. Pre-COVID, models often used supervised learning methods like Logistic Regression, Random Forests, or SVMs, calibrated on consistent health variables. Post-COVID, however, there’s a need for dynamic, adaptable models that consider new risk factors using ensemble or deep learning techniques to capture complex interactions between pre-existing conditions and post-viral impacts. COVID-19 has also accelerated the adoption of wearable devices and telemedicine, producing real-time data on heart rate variability, blood oxygen, and respiratory rates, which can enhance stroke prediction accuracy.

**1.2 Proposed System:**

The proposed system for Comparative Heart Stroke Prediction leverages machine learning to assess and predict stroke risks by integrating insights from both pre- and post-COVID-19 datasets. It is designed to identify and adapt to the nuanced risk factors introduced by COVID-19 while retaining the foundational cardiovascular health indicators traditionally used in stroke prediction. The system's data collection module aggregates information from electronic health records (EHRs), wearable health devices, and IoT-based monitors, incorporating both standard risk factors (e.g., age, blood pressure, cholesterol levels) and COVID-specific data (e.g., infection history, vaccine status, post-COVID symptoms like reduced lung function and blood oxygen levels). Data preprocessing and feature engineering steps are tailored to clean, normalize, and structure data from these varied sources, with a focus on engineering new COVID-relevant features. Machine learning models are then applied in a comparative framework, with separate models developed for pre- and post-COVID-19 data and later integrated to enhance predictive accuracy. Advanced models like ensemble methods and neural networks are trained to capture complex interactions between COVID-related factors and traditional stroke indicators. This comparative system ultimately provides healthcare providers with a robust tool for heart stroke risk assessment, helping them identify high-risk patients more accurately in the post-COVID landscape.

**1.3 Aim and Objective:**

**Aim:** The aim on “Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19” is to enhance the accuracy and relevance of heart stroke prediction models by examining and incorporating changes in cardiovascular risk factors influenced by the COVID-19 pandemic. Through a comparative analysis of machine learning insights from pre- and post-COVID-19 periods, this research seeks to better understand and adapt to any shifts in risk factors, thus improving predictive outcomes for diverse populations.

**Objectives**

1. **Identify Changes in Risk Factors**: Examine if traditional cardiovascular risk factors have altered in prevalence or impact due to COVID-19 and identify any new risk factors associated with the pandemic.
2. **Model Adaptation and Retraining**: Adapt and retrain existing machine learning models to reflect post-pandemic risk patterns, enhancing predictive accuracy and relevance.
3. **Address Demographic Disparities**: Assess and mitigate any potential biases in prediction models that may arise from COVID-19’s unequal impact on various demographic groups, ensuring fair and equitable predictions.
4. **Support Personalized Interventions**: Use refined predictive insights to guide personalized, preventative care for high-risk individuals, especially those impacted by pandemic-related cardiovascular health changes.
5. **Inform Public Health Policy**: Provide data-driven insights to assist in public health policy development, with a focus on resource allocation, awareness, and management of cardiovascular health risks in a post-COVID-19 landscape.

**1.4 Scope:**

The project's scope involves the study investigates heart stroke prediction using machine learning, focusing on changes in cardiovascular risk patterns due to the COVID-19 pandemic. It involves collecting and integrating health data from pre- and post-pandemic periods, including biometric information and lifestyle factors, to create comprehensive datasets. Various machine learning algorithms will be applied to develop predictive models that highlight shifts in traditional risk factors like hypertension and diabetes, as well as identify new risk factors that may have emerged due to the pandemic.

**CHAPTER-2**

## 2. Literature Survey

### Title: - Machine Learning and the Conundrum of Stroke Risk Prediction

**Author: -** Yaacoub Chahine, Matthew J Magoon, Bahetihazi Maidu, Juan C del Alamo, Patrick M Boyle, Nazem Akoum.

**Abstract:** Stroke is a leading cause of death worldwide. With escalating healthcare costs, early non-invasive stroke risk stratification is vital. The current paradigm of stroke risk assessment and mitigation is focused on clinical risk factors and comorbidities. Standard algorithms predict risk using regression-based statistical associations, which, while useful and easy to use, have moderate predictive accuracy. This review summarises recent efforts to deploy machine learning (ML) to predict stroke risk and enrich the understanding of the mechanisms underlying stroke. The surveyed body of literature includes studies comparing ML algorithms with conventional statistical models for predicting cardiovascular disease and, in particular, different stroke subtypes. Another avenue of research explored is ML as a means of enriching multiscale computational modelling, which holds great promise for revealing thrombogenesis mechanisms. Overall, ML offers a new approach to stroke risk stratification that accounts for subtle physiologic variants between patients, potentially leading to more reliable and personalised predictions than standard regression-based statistical associations.

**Keywords:** - Cardiovascular disease, computational modelling, neural networks, atrial fibrillation, thromboembolism, computational fluid dynamics, multiscale modelling.

.

### 

### Title: - A Review of Machine Learning’s Role in Cardiovascular Disease Prediction: Recent Advances and Future Challenges

**Author: -** Marwah Abdulrazzaq Naser, Aso Ahmed Majeed, Muntadher Alsabah, Taha Raad AI-Shaikhli, Kawa M.Kaky

**Abstract: -** Cardiovascular disease is the leading cause of global mortality and responsible for millions of deaths annually. The mortality rate and overall consequences of cardiac disease can be reduced with early disease detection. However, conventional diagnostic methods encounter various challenges, including delayed treatment and misdiagnoses, which can impede the course of treatment and raise healthcare costs. The application of artificial intelligence (AI) techniques, especially machine learning (ML) algorithms, offers a promising pathway to address these challenges. This paper emphasizes the central role of machine learning in cardiac health and focuses on precise cardiovascular disease prediction. In particular, this paper is driven by the urgent need to fully utilize the potential of machine learning to enhance cardiovascular disease prediction. In light of the continued progress in machine learning and the growing public health implications of cardiovascular disease, this paper aims to offer a comprehensive analysis of the topic. This review paper encompasses a wide range of topics, including the types of cardiovascular disease, the significance of machine learning, feature selection, the evaluation of machine learning models, data collection & preprocessing, evaluation metrics for cardiovascular disease prediction, and the recent trends & suggestion for future works.

**keywords:** Machine learning, cardiovascular disease, cardiovascular disease types; classification; prediction; cardiac care; feature selection; healthcare.

### Title: - An Interpretable Approach with Explainable AI for Heart Stroke Prediction

**Author: -** Parvathaneni Naga Srinivas, Uddagiri Sirisha, Kotte Sandeep, S. Phani Praveen, Lakshmana Phaneendra Maguluri, and Thulasi Bikku

**Abstract: -** Heart strokes are a significant global health concern, profoundly affecting the wellbeing of the population. Many research endeavors have focused on developing predictive models for heart strokes using ML and DL techniques. Nevertheless, prior studies have often failed to bridge the gap between complex ML models and their interpretability in clinical contexts, leaving healthcare professionals hesitant to embrace them for critical decision-making. This research introduces a meticulously designed, effective, and easily interpretable approach for heart stroke prediction, empowered by explainable AI techniques. Our contributions include a meticulously designed model, incorporating pivotal techniques such as resampling, data leakage prevention, feature selection, and emphasizing the model’s comprehensibility for healthcare practitioners. This multifaceted approach holds the potential to significantly impact the field of healthcare by offering a reliable and understandable tool for heart stroke prediction. In our research, we harnessed the potential of the Stroke Prediction Dataset, a valuable resource containing 11 distinct attributes. Applying these techniques, including model interpretability measures such as permutation importance and explainability methods like LIME, has achieved impressive results. While permutation importance provides insights into feature importance globally, LIME complements this by offering local and instance-specific explanations. Together, they contribute to a comprehensive understanding of the Artificial Neural Network (ANN) model.

**keywords:-** Artificial Neural Network, deep learning, data leakage, sampling, feature selection, explainable AI, LIME tabular.

### Title:- Stroke Risk Factor Prediction Using Machine Learning Techniques: A Systematic Review

**Author: -** Olusola Olabanjo, Ashiribo Wusu, Oseni Afisi and Boluwaji Akinnuwesi

**Abstract: -** This review addresses the global challenge of stroke, a leading cause of disability and mortality. The unpredictability and severe impact of stroke necessitate advanced prediction methods. In this work, the machine learning (ML) and deep learning (DL) techniques in stroke risk prediction were evaluated, assessing their effectiveness and application in diverse contexts. A systematic analysis of existing studies and datasets was conducted using Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA), focusing on various ML and DL algorithms used in stroke risk prediction. The 31 papers met the final inclusion criteria. The review highlights significant advancements in stroke prediction using ML and DL models, noting their ability to manage complex datasets and provide accurate predictions. However, challenges such as the need for external validation, model explainability and model transparency persist. Feature importance is further recommended to offer context-specific recommendations as stroke risk factors vary in different countries. This study also spotlights Random Forest as the outperforming model in predicting stroke risks, secondary data as the prominent dataset and China, India and Bangladesh as the country with the most stroke risk studies. The ML and DL offer promising tools for stroke risk prediction, enhancing personalized healthcare strategies. Addressing existing challenges will be crucial for their effective integration into clinical practice.

**keywords:**- Stroke risk prediction, machine learning, deep learning, PRISMA review, predictive models, personalized medicine

#### Title: - Machine learning-based classification of valvular heart disease using cardiovascular risk factors

**Author: -** Muhammad Usman Aslam, Songhua Xu, Sajid Hussain, Muhammad Waqas and Nafiu

Lukman Abiodun

**Abstract:-** Valvular Heart Disease (VHD) is a globally significant cause of mortality, particularly among aging populations. Despite advancements in percutaneous and surgical interventions, there are still uncertainties that remain regarding the risk factors that significantly contribute to this condition within the domain of cardiovascular disease. This study investigates these uncertainties and the role of machine learning in categorizing VHD based on cardiovascular risk factors. It follows a two-part investigation comprising feature extraction and classification phases. Feature extraction is initially performed using a wrapping approach and refined further with binary logistic regression. The second phase employs five classifiers: Artificial Neural Network (ANN), XGBoost, Random Forest (RF), Naïve Bayes, and Support Vector Machine (SVM), along with advanced methods such as SVM combined with Principal Component Analysis (PCA) and a majority-voting ensemble method (MV5). Data on VHD cases were collected from DHQ Hospital Faisalabad using simple random sampling. Various statistical measures, such as the ROC curve, F-measure, sensitivity, specificity, accuracy, MCC, and Kappa are applied to assess the results. The findings reveal that the combination of SVM with PCA achieves the highest overall performance while the MV5 ensemble method also demonstrates high accuracy and balance in sensitivity and specificity. The variation in VHD prevalence linked to specific risk factors highlights the importance of a comprehensive approach to reduce this disease’s burden.

# CHAPTER – 3

## 3. Design

* 1. **Software Requirements:**
  2. Python is a leading choice for machine learning in healthcare due to its vast ecosystem of libraries, tools, and active community support. It allows efficient data handling, processing, and advanced statistical and predictive analysis.
  3. NumPy provide reliable data manipulation capabilities, supporting efficient transformation of raw data into formats suitable for machine learning.
  4. SciPy is also essential for statistical functions, such as hypothesis testing and feature analysis, ensuring that models are built on robust, meaningful variables.
  5. Matplotlib and Seaborn are essential for basic EDA (Exploratory Data Analysis), while Plotly or Power BI can be useful for creating interactive visuals, particularly helpful in presentations or collaboration with stakeholders.
  6. R is also an option, especially when detailed statistical testing and data visualization are priorities.
  7. TensorFlow or PyTorch provide frameworks for advanced neural networks.
  8. Anaconda helps manage libraries and package dependencies, which is crucial when working across different development environments.
  9. Docker can containerize the ML environment, ensuring consistency across various machines and aiding in deployment.

### 3.2 Functional Requirements

The functional requirements for a comparative heart stroke prediction system using machine learning, particularly one that incorporates pre- and post-COVID-19 insights, include detailed functionality from data management to model deployment and interpretability.

**3.2.1** **Data Collection and Ingestion**

* **Data Integration:** Collect historical data on cardiovascular risk factors (e.g., blood pressure, cholesterol levels, smoking habits) and COVID-19 related factors (e.g., infection severity, hospitalization details) from various sources like electronic health records (EHR), clinical trials, and epidemiological studies.
* **Real-Time Data Updates:** Enable automated ingestion of new patient data in real-time, especially from EHRs, to keep predictions current with recent cases and treatments.

**3.2.2 Data Processing and Feature Engineering**

* **Feature Extraction:** Extract relevant features for stroke prediction, such as age, BMI, history of hypertension, COVID-19 severity, recovery duration, and respiratory complications.
* **Feature Transformation:** Normalize or standardize continuous features (e.g., age, BMI) and encode categorical variables (e.g., COVID-19 severity as mild, moderate, severe).
* **New Feature Creation:** Engineer new features specific to COVID-19 impacts, such as “COVID-19 Complications Score” based on hospitalization severity or ICU admission history.

**3.2.3** **Model Selection and Training**

* **Model Training and Validation:** Train machine learning models such as logistic regression, random forest, and neural networks, evaluating each on historical (pre-COVID-19) data and recent (post-COVID-19) data.
* **Algorithm Selection:** Provide options to choose the best-suited model based on evaluation metrics for pre-COVID-19 and post-COVID-19 datasets.
* **Cross-Validation:** Implement cross-validation techniques to avoid overfitting and ensure the model generalizes well across different patient populations.

**3.2.4 Model Evaluation and Validation:**

* **Performance Metrics Calculation**: Calculate and display key evaluation metrics (e.g., Precision, Recall, F1-score, ROC-AUC) for both pre- and post-COVID-19 models to facilitate comparison.
* **Bias and Fairness Assessment:** Assess model fairness, ensuring no significant bias across demographic groups (e.g., age, gender, ethnicity) to prevent discrimination.
* **Model Drift Monitoring**: Continuously monitor for concept drift due to changing risk factors or COVID-19 complications and update models as necessary.

**3.2.5 Prediction and Risk Scoring**

* **Risk Score Calculation:** Provide a clear risk score indicating the likelihood of heart stroke, adjusted for pre- and post-COVID-19 data nuances.
* **Threshold Customization:** Allow medical professionals to set custom threshold levels for risk scores, adjusting sensitivity according to their clinical practice.
* **Prediction with Explanation:** Deliver predictions with explainability tools (e.g., SHAP or LIME), showing factors contributing most to each patient’s stroke risk.

**3.2.6 Patient Dashboard and Visualization:**

* **Patient Profile Dashboard**: Show a dashboard with patient details, including past and current risk factors, COVID-19-related data, and predicted risk score.
* **Comparison Visualization:** Present pre- and post-COVID-19 risk predictions side-by-side, along with graphical representations (e.g., bar charts, trend lines) to highlight changes.
* **Interactive Tools**: Provide tools for clinicians to explore how changes in patient data (e.g., improved BMI, recovery from COVID-19) impact risk scores dynamically.

**3.2.7 Model Deployment and Integration:**

* **API Development**: Develop APIs for model predictions, enabling integration with healthcare systems and electronic health records.
* **Cloud or On-Premise Deployment**: Support deployment in either cloud or on-premise environments, depending on the healthcare provider's infrastructure and data privacy requirements.
* **Batch and Real-Time Prediction Modes**: Allow for both batch prediction (for large datasets) and real-time prediction for individual patients during medical consultations.

**3.2.8** **Explainability and Interpretability:**

* **Feature Importance Ranking**: Display rankings of influential features contributing to stroke prediction, aiding clinicians in understanding model reasoning.
* **Explainability Tools**: Integrate SHAP or LIME to provide explanations for each prediction, enhancing trust in the model’s decision-making process.
* **Post-Hoc Analysis**: Allow clinicians to conduct analysis on how COVID-19-specific features (e.g., severity, recovery duration) are influencing stroke predictions, supporting data-driven decisions.

**3.2.9 User Role and Access Management:**

* **Role-Based Access Control**: Enable different access levels for clinicians, data scientists, and administrators, ensuring patient data privacy and security.
* **Audit Logs:** Maintain a log of data access and model prediction requests to monitor usage and ensure compliance with data protection regulations like HIPAA and GDPR.

**3.2.10 Compliance and Security:**

* **Data Encryption**: Ensure all patient data is encrypted in transit and at rest to comply with healthcare data privacy standards.
* **Anonymization and De-identification**: Anonymize or de-identify patient information where necessary, especially when sharing data for model training or collaborative studies.
* **Compliance Adherence:** Ensure that all software components meet relevant regulatory standards, including HIPAA, GDPR, and local health data laws.

**3.2.11 Continuous Improvement and Updates:**

* **Model Retraining Pipeline:** Establish a retraining pipeline to update the model periodically with new data, particularly to reflect changing post-COVID-19 trends.
* **Feedback Loop:** Allow clinicians to provide feedback on predictions, improving model performance and interpretability over time.
* **Performance Monitoring:** Continuously monitor prediction accuracy and revalidate models, alerting for potential retraining as data distributions change.

### 3.3 Non-Functional Requirements

The non-functional requirements for a comparative heart stroke prediction system using machine learning, addressing insights pre- and post-COVID-19, are essential for system reliability, performance, and user trust. Here’s a breakdown of the key non-functional requirements:

**3.3.1. Performance:**

* **Response Time**: The system should deliver predictions and risk scores within a few seconds (preferably under 5 seconds) to allow real-time usage during medical consultations.
* **Throughput**: The system should support batch processing of multiple patient records, handling at least hundreds of predictions per minute for large-scale analysis.
* **Scalability**: The architecture must support horizontal scaling to handle increased data loads, especially during peak usage or high-volume healthcare facilities.

**3.3.2. Reliability and Availability:**

* **System Uptime:** The system should maintain 99.9% availability to ensure constant access for healthcare professionals.
* **Error Handling:** The system must provide detailed error logging and user-friendly messages in case of failures, ensuring the user is informed without disrupting critical operations.
  + **Automatic Recovery:** Implement automatic recovery mechanisms for data ingestion, prediction, and storage in case of system failures or network issues.

**3.3.3 Accuracy and Precision**

* **Prediction Accuracy**: The system must meet a minimum predictive accuracy (e.g., ROC-AUC score > 0.85) across both pre- and post-COVID-19 data to ensure clinical utility.
* **Consistency:** Predictions should be consistent across repeated requests, ensuring stability in risk scores for the same patient under identical conditions.
  + **Data Validation:** Include mechanisms to validate input data format and structure before processing, minimizing errors due to data inconsistencies.

**3.3.4 Usability:**

* **User Interface Simplicity**: Ensure an intuitive, streamlined UI with minimal navigation steps, accessible even to users with basic technical skills.
* **Visualization Clarity:** Data visualizations must be clear and customizable, making it easy for clinicians to interpret risk scores and key features influencing predictions.
* **Mobile and Desktop Compatibility:** The system should be compatible with both desktop and mobile devices, accommodating diverse clinical settings.

**3.3.5 Scalability**

* **Data Storage Scalability:** The system must accommodate growing data volumes as more patients’ records are added, especially in large healthcare organizations.
* **Model Scalability:** The infrastructure should support the deployment of multiple models if needed, enabling comparison of multiple models for pre- and post-COVID-19 data.
* **Resource Elasticity:** Use cloud or on-premise elastic resources that adjust based on demand, allowing cost-effective operation without compromising performance.

**3.3.6 Security**

* **Data Encryption:** Ensure end-to-end encryption for patient data both at rest and in transit, aligning with healthcare security standards like HIPAA and GDPR.
* **Role-Based Access Control:** Implement role-based access management to restrict data access, protecting sensitive patient information.
* **Audit Trails**: Maintain comprehensive logs for all access and prediction requests to monitor system use and detect unauthorized access.

**3.3.7** **Interoperability**

* **Integration with EHR Systems:** Support integration with standard electronic health record (EHR) systems (e.g., HL7, FHIR) for seamless data transfer and access.
* **API Compatibility:** Develop API endpoints following industry standards to facilitate integration with third-party applications or other healthcare systems.
* **Data Standardization**: Use standardized medical terminologies (e.g., ICD, SNOMED) to ensure consistent data representation across systems and improve data interpretability.

**3.3.8 Maintainability**

* **Modular Codebase**: Develop a modular architecture to allow individual components (e.g., data ingestion, prediction model) to be updated independently without affecting the entire system.
* **Continuous Integration and Continuous Deployment (CI/CD):** Implement a CI/CD pipeline for smooth deployment of updates, allowing frequent model and software updates without downtime.
  + **Documentation:** Provide detailed, well-organized documentation for the system's architecture, API endpoints, and deployment processes to facilitate maintenance and onboarding.

**3.3.9** **Explainability and Transparency**

* **Model Interpretability**: Ensure that clinicians can understand the model’s decision-making process through explainability tools like SHAP and LIME, making it transparent which factors influence stroke risk.
* **Traceability of Predictions:** Each prediction should be traceable to the specific data inputs and model versions used, ensuring transparency in how risk scores are calculated.
* **Model Versioning:** Maintain and display model version information to ensure users are aware of the specific model in use, especially when performance varies between pre- and post-COVID-19 data.

**3.3.10. Compliance**

* **Regulatory Compliance:** The system must adhere to healthcare regulations (e.g., HIPAA in the U.S., GDPR in Europe) to ensure legal compliance in handling patient data.
* **Data Residency:** Ensure compliance with local data residency laws, storing patient data within geographic boundaries as mandated by regional regulations.
* **Audit and Reporting Features:** Provide compliance reporting capabilities, ensuring all data access and processing activities are traceable and auditable by regulatory bodies.

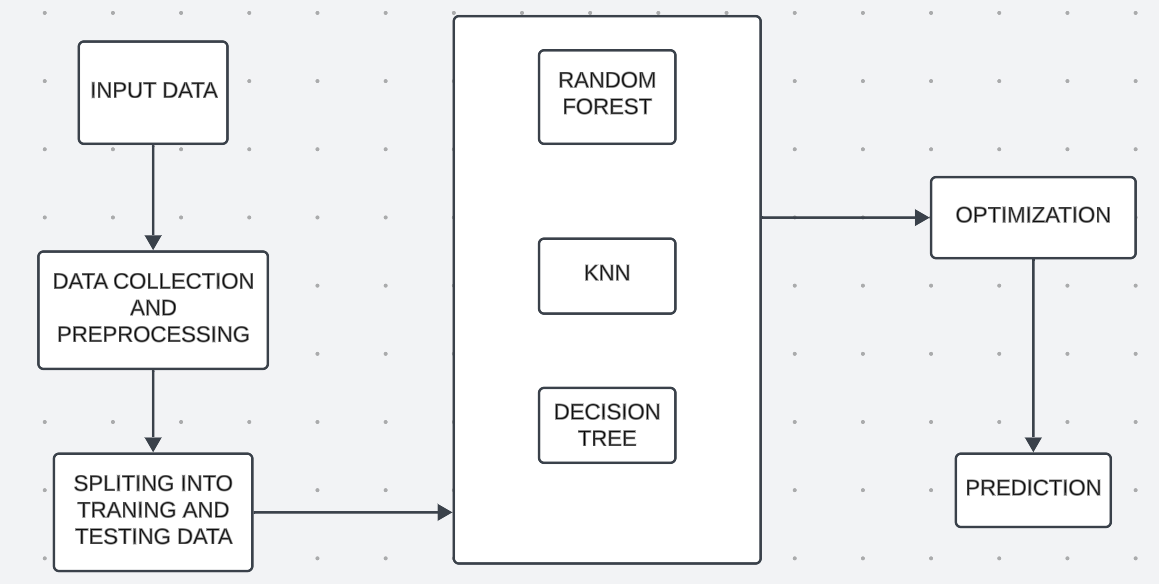
**3.3.11 Adaptability**

* **Model Update Flexibility**: Enable rapid deployment of new models as medical research on COVID-19 evolves, ensuring the system adapts to updated risk factors and prediction models.
* **Feature Adjustment:** Allow dynamic addition or modification of input features as new COVID-19 variants or post-recovery conditions are identified.
* **Continuous Learning:** Incorporate a feedback mechanism for clinicians to submit feedback on predictions, which can be used to improve model accuracy over time.

**3.3.12 Cost-Efficiency**

* **Resource Optimization**: Optimize computational and storage resources, balancing cost with performance, especially for large-scale healthcare applications.
* **Elastic Cloud Resources**: Use cloud-based elastic resources that scale up or down as needed, reducing costs during periods of low usage.
* **Licensing Flexibility**: Ensure that any third-party software or libraries used allow for cost-effective licensing, supporting large-scale deployment without excessive cost.

**3.4 Model Architecture:**



#### Fig 3.4: System Architecture

# CHAPTER-4

## Implementation

For implementing a comparative heart stroke prediction study that leverages machine learning insights pre- and post-COVID-19, here’s a structured approach you can follow:

**4.1 Data Acquisition and Preparation**

* **Data Sources:** Obtain relevant datasets from sources like hospital EHRs, the CDC, WHO, or public health research repositories. Ensure the data contains timestamps, patient demographics, stroke risk factors (e.g., hypertension, diabetes, obesity), and COVID-19-related information (e.g., infection history, vaccination status).
* **Data Cleaning:** Handle missing values (e.g., through imputation or deletion), remove duplicates, and address inconsistencies.
* **Feature Engineering:** Createfeatures representing traditional stroke risk factors and new COVID-19-related features like:

 **Demographics:** Age, sex, and race.

 **Clinical Indicators:** Hypertension, diabetes, cholesterol levels, heart disease, smoking status.

 **COVID-19 Specifics:** COVID-19 infection history, severity, long-COVID symptoms, vaccination status, and hospitalizations.

* **Data Splitting:** Divide the dataset into pre-COVID-19 and post-COVID-19 groups based on dates. Use the same preprocessing steps for both datasets to maintain consistency.

**4.2 Exploratory Data Analysis (EDA)**

* **Trend Analysis:** Assess how risk factors and stroke incidence rates have changed pre- and post-COVID-19.
* **Visualization:** Use box plots, histograms, and correlation matrices to observe the distributions and relationships of the variables.
* **Statistical Testing:** Perform t-tests, ANOVA, or chi-square tests to determine if there are significant changes in feature distributions or outcomes post-COVID-19.

**4.3 Model Selection and Training**

* **Model Choices:** Select a range of machine learning models for stroke prediction, such as:
* **Logistic Regression:** For baseline comparison and understanding linear relationships.
* **Random Forest and Gradient Boosting:** For handling non-linear relationships and feature importance analysis.
* **K-Nearest Neighbors (KNN):** For high-dimensional feature spaces.
* **Neural Networks:** Particularly useful if you have a large dataset with complex patterns.
* **Cross-Validation:** Use k-fold cross-validation to improve model reliability and mitigate overfitting.
* **Hyperparameter Tuning:** Employ grid search or random search techniques for optimizing model parameters.

**4.4 Training and Evaluation on Pre- and Post-COVID-19 Data**

* **Separate Training**: Train the models on pre-COVID-19 and post-COVID-19 datasets independently to detect any shifts in prediction patterns or accuracy.
* **Evaluation Metrics**: Track and compare metrics like accuracy, precision, recall, F1-score, and AUC-ROC across both datasets to evaluate performance. If post-COVID-19 data introduces new complexities, it may result in lower model accuracy or different feature importance.
* **Feature Importance Analysis**: Use methods like SHAP (SHapley Additive exPlanations) to compare how the importance of each feature changes pre- and post-COVID-19.

**4.5 Comparative Analysis and Insights**

* **Performance Comparison**: Analyze differences in performance metrics for the models trained on pre- and post-COVID-19 data to understand changes in predictability.
* **Feature Importance Shift**: Identify whether COVID-19-related variables (e.g., infection history, vaccination status) significantly affect stroke prediction in post-COVID-19 data.
* **Risk Factor Analysis**: Use insights from the models to highlight new or intensified risk factors in the post-COVID-19 data. For instance, long-COVID symptoms might emerge as a new factor influencing stroke risk.

**4.7 Review and Adjustment of Access Rights**

As the user becomes more familiar with the system and as operational needs evolve, it may be necessary to review and adjust access rights and permissions. This ensures that the user remains equipped with the tools they need to manage their area effectively while maintaining system security. Periodic assessments of user roles and responsibilities can help determine if additional permissions are required or if any changes are necessary.

**4.8 Algorithms**

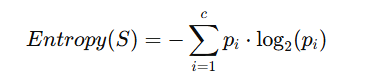
**4.8.1 Random Forest**

Random forests are ensemble models that use multiple decision trees to capture non-linear relationships within the data. They are particularly useful for stroke prediction because they handle a mix of numerical and categorical variables well and can identify important predictors, which is valuable in understanding how COVID-19-related factors impact stroke risk. Random forests also provide a degree of interpretability through feature importance scores, enabling a comparative analysis of risk factors pre- and post-pandemic.

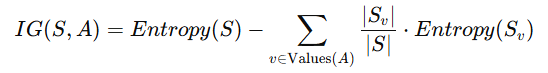
**4.8.2 Decision Tree**

The decision tree algorithm is a pivotal machine learning technique used in the study "Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19" to analyze patterns in patient data and predict stroke risks effectively. A decision tree functions by recursively splitting the dataset based on the values of specific features, creating a tree-like structure where each internal node represents a decision rule, branches represent possible outcomes, and leaf nodes correspond to predictions.

This algorithm employs metrics like Information Gain (IG) and Entropy to identify the most informative splits. Entropy, a measure of disorder in the data, is defined as:



where S is the dataset, pi is the proportion of instances in class iii, and ccc is the number of classes. Information Gain quantifies the reduction in entropy achieved by splitting on a feature A, calculated as:



where Sv​ is the subset of S where feature A has value v and ∣Sv∣/∣S∣ is the proportion of samples in Sv​.

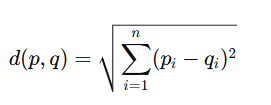
In the study, decision trees enabled the identification of critical factors affecting stroke risk, such as age, comorbidities, and lifestyle changes. By comparing results pre- and post-COVID-19, the algorithm highlighted shifts in these factors, reflecting the pandemic's impact on healthcare access and patient behavior. This approach provided interpretable insights, aiding in the development of targeted interventions and improving stroke prediction accuracy during different timeframes.

### 4.8.3 Support Vector Machine (SVM)

Support vector machines are effective for high-dimensional feature spaces and are useful if the dataset has many features or noise. SVMs work by finding the optimal hyperplane that maximally separates the stroke and non-stroke classes. While less interpretable, they can reveal nuanced changes in feature separability pre- and post-COVID-19 and are worth testing for their robustness in distinguishing between patients at higher risk of stroke.

### 4.8.4 K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm is a simple yet powerful machine learning method for heart stroke prediction. It classifies a new data point by considering the majority class among its k nearest neighbors in the feature space. The distance between data points is calculated using a metric such as Euclidean distance, defined as:



where p and q are two points in an n-dimensional space, and pi and qi represent the values of the i-th feature of p and q, respectively. For heart stroke prediction, the algorithm would analyze features such as age, blood pressure, cholesterol levels, and comorbidities. By comparing predictions pre- and post-COVID-19, variations in key risk factors, such as the influence of COVID-related health changes, can be quantitatively assessed. Adjusting the value of k ensures that the algorithm balances between sensitivity to local noise (k too small) and over smoothing (k too large), optimizing stroke prediction accuracy.

### 4.8.5 Artificial Neural Networks (ANN)

### Neural networks, especially deep learning models, are suited for large, complex datasets and can capture non-linear relationships between features. ANN can be particularly effective in discovering intricate relationships among risk factors that might not be obvious in simpler models. For stroke prediction, ANN can reveal complex interactions between traditional and COVID-19-specific features, although they are less interpretable compared to other algorithms and typically require more computational power.

### 4.9 Algorithm Comparison and Selection

### After training these models on both pre- and post-COVID-19 datasets, compare their performance metrics such as accuracy, F1-score, precision, recall, and AUC-ROC. Evaluate how each model captures risk factors differently in each dataset, particularly any COVID-19-related features, and determine which model generalizes well across both datasets.

### This suite of algorithms provides a comprehensive approach to capturing and comparing stroke prediction patterns, enabling a deeper understanding of the pandemic's impact on cardiovascular health.

### 4.10 Libraries

* 1. **Pandas:** The core library for data manipulation and analysis. Use Pandas to load, clean, and preprocess data, as well as to perform feature engineering. Its DataFrame structure is ideal for handling large datasets.
  2. **NumPy**: Essential for numerical operations, especially for array manipulations and computations. It’s often used alongside Pandas for efficient handling of numerical data.
  3. **Matplotlib** and **Seaborn**: Both libraries are excellent for creating visualizations to explore data distributions, trends, and relationships. Matplotlib provides low-level plotting functions, while Seaborn simplifies creating statistical plots like histograms, box plots, and correlation heatmaps.

**Main.py for Pre Covid-19**

import numpy as np

import pandas as pd

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

import os

import io

df = pd.read\_csv(io.BytesIO(uploaded['healthcare-dataset-stroke-data.csv']))

df.head()

df.shape

df.describe().T

df.isnull().sum()

df.interpolate(inplace=True)

df.fillna(method='ffill',inplace=True)

df.fillna(method='bfill',inplace=True)

df.isnull().sum()

sns.countplot(df["smoking\_status"])

Df=pd.get\_dummies(df,columns=["gender","work\_type","smoking\_status"])

Df.head()

Df.drop('gender\_Female',axis=1,inplace=True)

Df.drop('work\_type\_children',axis=1,inplace=True)

Df.drop('smoking\_status\_smokes',axis=1,inplace=True)

Df.head()

from sklearn.tree import DecisionTreeClassifier

from sklearn.metrics import accuracy\_score,precision\_score,recall\_score,f1\_score

from sklearn.model\_selection import train\_test\_split

model=DecisionTreeClassifier()

X=Df[["age","hypertension","heart\_disease","avg\_glucose\_level","bmi","gender\_Male","gender\_Other","work\_type\_Govt\_job", "work\_type\_Never\_worked","work\_type\_Private","work\_type\_Self-employed", "smoking\_status\_formerly smoked","smoking\_status\_never smoked"]]

y=Df["stroke"]

X\_train,X\_test,y\_train,y\_test = train\_test\_split(X,y,test\_size=0.3,random\_state=2)

model.fit(X\_train,y\_train)

y\_pred = model.predict(X\_test)

print("accuracy score is ",accuracy\_score(y\_pred,y\_test))

print("precision score is ",precision\_score(y\_pred,y\_test))

print("recall score is ",recall\_score(y\_pred,y\_test))

print("f1 score is ",f1\_score(y\_pred,y\_test))

sns.countplot(df["smoking\_status"])

sns.scatterplot(data=df, x='avg\_glucose\_level', y='bmi', hue='gender')

import matplotlib.pyplot as plt

from sklearn.metrics import classification\_report, confusion\_matrix, accuracy\_score

import seaborn as sns

plt.figure(figsize=(8, 6))

sns.heatmap(confusion\_matrix(y\_test, y\_pred), annot=True, fmt='d', cmap='Blues')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix')

plt.show()

**Main.py for Post Covid-19**

!pip install pandas scikit-learn matplotlib seaborn

!pip install pandas scikit-learn

import pandas as pd

import io

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import StandardScaler

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, confusion\_matrix, accuracy\_score

import matplotlib.pyplot as plt

import seaborn as sns

# Step 1: Load the dataset

  # Update this with your actual file path

data = pd.read\_csv(io.BytesIO(uploaded['dataset.csv']))

# Step 2: Explore the data

print(data.head())

print(data.info())

print(data.describe())

# Step 3: Preprocess the data

# Convert categorical variables if necessary (gender, cp, restecg, slope, thal)

data['sex'] = data['sex'].replace({0: 'Female', 1: 'Male'})

# data['sex'] = data['sex'].replace({0: 'Female', 1: 'Male'}).astype('category')

data['cp'] = data['cp'].astype('category')

data['restecg'] = data['restecg'].astype('category')

data['slope'] = data['slope'].astype('category')

data['thal'] = data['thal'].astype('category')

print(data.head())

# One-hot encode categorical variables

data = pd.get\_dummies(data, drop\_first=True)

# Step 4: Split the data into features and target

X = data.drop('target', axis=1)  # Features

y = data['target']  # Target

# Step 5: Split the data into training and testing sets

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

# Step 6: Standardize the features

scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

# Step 7: Create and train the model

model = RandomForestClassifier(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

# Step 8: Make predictions

y\_pred = model.predict(X\_test)

# Step 9: Evaluate the model

print("Confusion Matrix:")

print(confusion\_matrix(y\_test, y\_pred))

print("\nClassification Report:")

print(classification\_report(y\_test, y\_pred))

print("\nAccuracy Score:", accuracy\_score(y\_test, y\_pred))

# Step 10: Visualize the confusion matrix

plt.figure(figsize=(8, 6))

sns.heatmap(confusion\_matrix(y\_test, y\_pred), annot=True, fmt='d', cmap='Blues')

plt.xlabel('Predicted')

plt.ylabel('Actual')

plt.title('Confusion Matrix')

plt.show()

data.hist(figsize=(10,10))

plt.suptitle('Health Data Distribution: Histogram Analysis', fontsize=16)

plt.show()

import matplotlib.pyplot as plt

for column in data.columns:

plt.figure(figsize=(10, 6))

data[column].value\_counts().plot(kind='bar', color='blue')

plt.title(f'Value Counts for {column}')

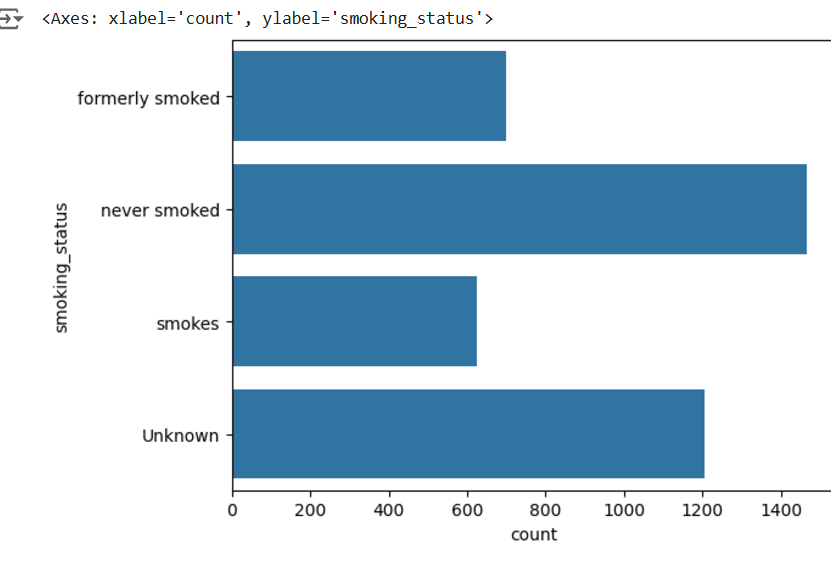
plt.xlabel(column)

plt.ylabel('Count')

plt.show()

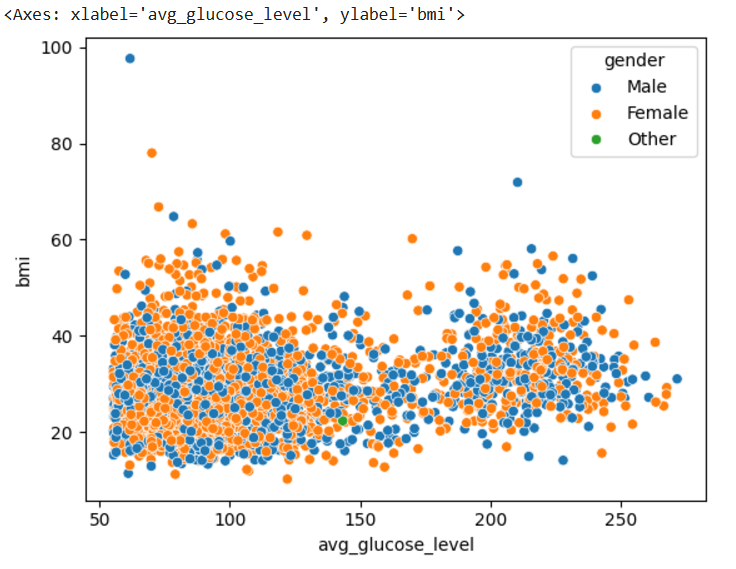
# CHAPTER-5

## 5. Results



**Fig 5.1 Smoking Status**

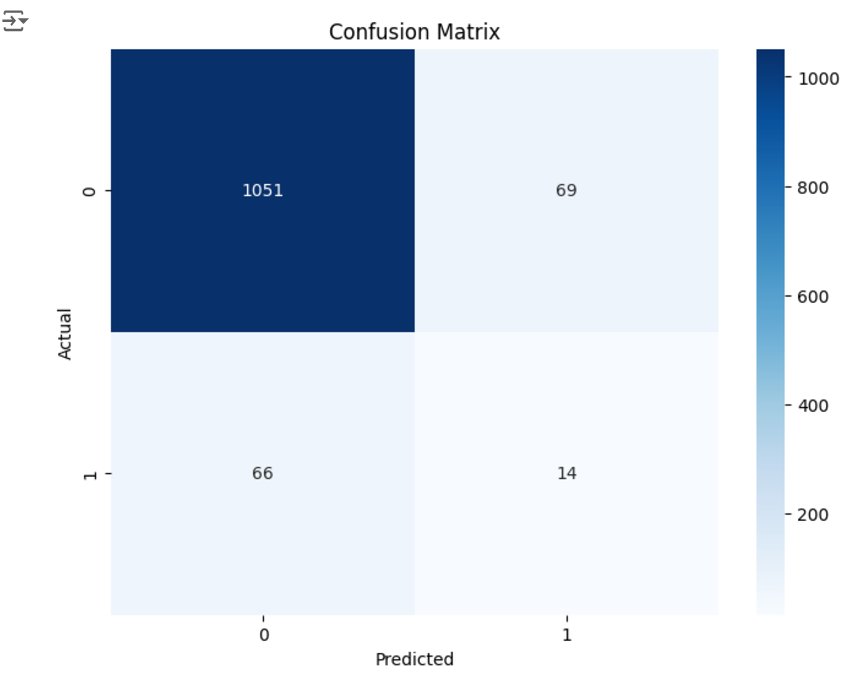
Fig 5.1: The bar chart shows the distribution of individuals by smoking status: "never smoked" has the highest count, followed by "Unknown," "formerly smoked," and "smokes" with the lowest count.



**Fig 5.2 Glucose levels**

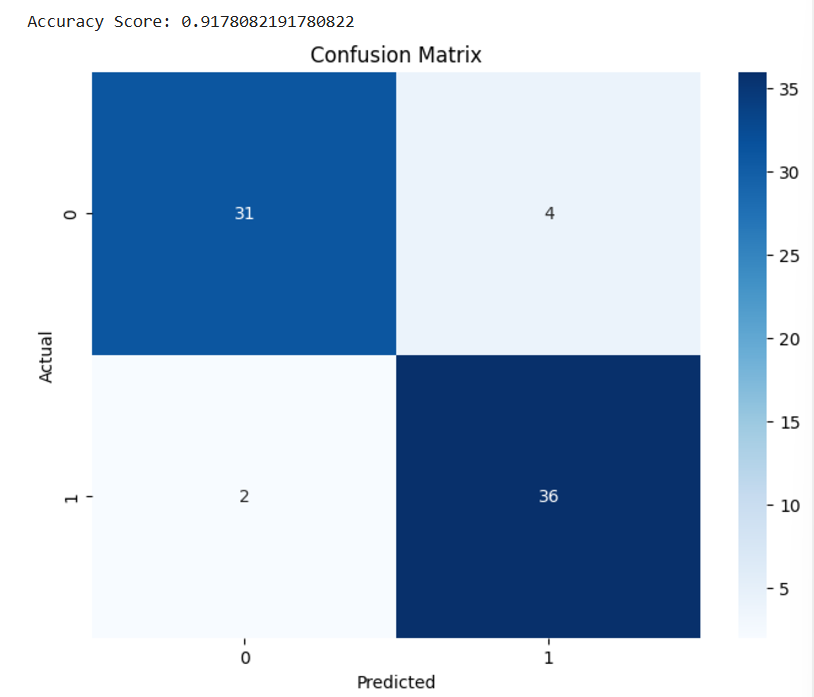
Fig 5.2: The scatter plot shows the relationship between average glucose level and BMI, categorized by gender (Male, Female, Other). BMI values cluster between 15 and 50 across all glucose levels.





**Fig.5.3 Output for Pre-Covid**

Fig 5.3: The Accuracy for this dataset is 0.88 and This figure explains 1051 people will not get heart stroke, 69 people will not get heart stroke, 66 people may get heart stroke, 14 people get heart stroke based on the data.



**5.4** **Output for Post-Covid**

Fig 5.4: The Accuracy for this dataset is 0.91 and This figure explains 31 people will not get heart stroke, 4 people will not get heart stroke, 2 people may get heart stroke, 36 people get heart stroke based on the data.

### 

# CHAPTER-6

## 6. Testing

**6.1 System Testing:**

### System testing for a comparative heart stroke prediction model involves thoroughly evaluating the model’s performance and accuracy across both pre- and post-COVID-19 datasets to ensure it is reliable and robust. This testing begins with unit tests to validate the functionality of individual components, such as data preprocessing, feature selection, and model training. Integration testing follows to confirm that these components work seamlessly together, especially with new COVID-19-related variables. Performance tests then assess the model’s accuracy, sensitivity, and specificity on both datasets, using metrics like AUC-ROC and F1-score to measure predictive power. Cross-validation techniques are employed to verify that the model generalizes well to diverse data and mitigates overfitting. Stress testing is crucial for models using real-time data from wearables, ensuring the system remains stable under high data volume or frequent updates. Finally, user acceptance testing, where clinicians review model outputs for interpretability and reliability, is essential for gaining trust and ensuring the system aligns with practical healthcare needs. Through these testing stages, the model is fine-tuned to be a responsive, accurate tool in stroke prediction both pre- and post-pandemic.

### 6.2 Software Testing Strategies:

Software testing strategies for a comparative heart stroke prediction system, which addresses both pre- and post-COVID-19 data, involve a layered approach to ensure accuracy, reliability, and adaptability. Unit testing begins the process by verifying each component, such as data preprocessing, feature extraction, and machine learning algorithms, to identify any early issues. Integration testing follows, checking that all components, including new COVID-19-specific variables, function cohesively to enable smooth data flow through the pipeline without errors or performance bottlenecks. System testing then evaluates the model’s performance across complete datasets from both periods, ensuring consistent and accurate stroke risk identification across different scenarios. Performance testing assesses the model’s speed, scalability, and resource efficiency, which are critical for handling large volumes of data, particularly if real-time inputs from wearable devices or telemedicine are included.

Further, regression testing is conducted to verify that the addition of COVID-19 health metrics or updates does not negatively affect existing functionality, ensuring the model remains consistent in its predictive accuracy over time. Finally, user acceptance testing is vital, as it involves feedback from clinicians and healthcare providers, ensuring the model's predictions are interpretable, actionable, and reliable in a real-world clinical setting. This thorough, multi-level testing process supports the creation of a robust, adaptable stroke prediction model that can respond effectively to the evolving healthcare landscape post-COVID-19. By addressing the complexities introduced by COVID-19-related health metrics, these testing strategies help create a predictive tool that remains accurate and practical in identifying stroke risks amidst new and traditional health variables.

**6.3 TESTING**

Testing for the comparative heart stroke prediction system, analyzing machine learning insights from pre- and post-COVID-19 datasets, involves a structured approach to ensure accuracy and reliability. It begins with unit testing to evaluate individual components like data preprocessing and algorithms, followed by integration testing to confirm seamless functionality of new COVID-19-specific features. System testing assesses overall performance on complete datasets, while performance testing ensures efficiency and scalability for real-time data processing. Regression testing checks that updates don’t compromise existing features, and user acceptance testing gathers clinician feedback on interpretability and practical use in clinical settings. This comprehensive strategy enhances the model's robustness and adaptability, supporting timely stroke risk assessments in the evolving healthcare landscape.

**6.3.1 TESTING METHODS**

Testing methods for comparative heart stroke prediction pre- and post-COVID-19, here’s a structured approach:

**6.3.1.1 Define the Baseline Models (Pre-COVID)**

Select Common Algorithms: Use traditional machine learning models such as Logistic Regression, Random Forest, SVM, and Neural Networks, which were popular pre-COVID.

Feature Set: Include features like age, blood pressure, cholesterol, diabetes, smoking, and lifestyle factors.

Performance Metrics: Evaluate these models using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to establish a baseline.

**6.3.1.2** **Introduce Post-COVID Risk Factors**

Data Augmentation: Integrate new risk factors related to COVID-19 complications, such as blood clotting markers (e.g., D-dimer), inflammation levels (e.g., C-reactive protein), and history of COVID infection.

Post-COVID Dataset Creation: If separate datasets (pre- and post-COVID) are available, treat them separately. Otherwise, label cases as "Pre" or "Post-COVID" if merged, allowing for comparative analysis.

**6.3.1.3 Model Adaptation and Training**

Adaptation of Baseline Models: Retrain the pre-COVID models on the post-COVID dataset, incorporating the new features to examine if they improve prediction accuracy.

Train New Models: Train separate models using only the post-COVID data to capture unique risk profiles.

Use Ensemble Techniques: Combine models trained on both pre- and post-COVID features to see if ensemble approaches yield better results.

**6.3.1.4 Performance Evaluation and Comparison**

Metrics Comparison: Compare the accuracy, precision, recall, F1-score, and ROC-AUC of pre-COVID models, post-COVID models, and hybrid models to analyze predictive performance differences.

Feature Importance Analysis: Use methods like SHAP (SHapley Additive exPlanations) or feature importance scores to identify the most significant features in pre- and post-COVID models.

Statistical Testing: Apply statistical tests like McNemar's test to determine if differences in model performance are significant.

**6.3.1.5. Interpretation of Results and Insights**

Pre- vs. Post-COVID Insights: Identify which features gained or lost significance post-COVID, providing insights into how COVID-19 has changed risk profiles for heart stroke.

Model Robustness: Evaluate how well the models generalize across datasets, which is essential for real-world applicability.

Guidance for Clinical Practice: Based on the findings, propose adjustments in clinical screenings or interventions for heart stroke in post-COVID patients.

# CHAPTER-7

## 7. Conclusion

The comparative analysis of heart stroke prediction models before and after COVID-19 reveals how the pandemic has reshaped cardiovascular risk assessment and machine learning applications in healthcare. Pre-COVID-19 models largely focused on traditional risk factors, but the emergence of new health challenges such as post-viral cardiovascular complications and inflammation has highlighted the need for more dynamic, adaptable models. Post-COVID-19 insights underscore the importance of integrating new data sources, including wearable health data, telemedicine inputs, and COVID-19-specific biomarkers, to improve predictive accuracy and responsiveness. Machine learning models equipped to handle these complex, evolving data landscapes can offer a more precise, personalized approach to stroke prediction, ensuring that healthcare systems remain responsive to the ongoing impacts of the pandemic. Through ongoing enhancements in technology and data integration, predictive models can better anticipate stroke risks, ultimately improving preventive care and patient outcomes in a post-COVID-19 world.

# CHAPTER-8

## 8. Future Enhancement

Future enhancements in comparative heart stroke prediction models can leverage advanced machine learning techniques, new data sources, and improved robustness to adapt to evolving health insights post-COVID-19. Integrating real-time health data from wearable devices such as heart rate, oxygen saturation, and sleep patterns along with telehealth data, like self-reported symptoms and medication adherence, can provide more precise, individualized stroke risk predictions. Advanced machine learning techniques, including Explainable AI (XAI) through methods like SHAP or LIME, can improve clinician trust by clarifying how individual risk factors influence predictions. Additionally, Auto ML frameworks can help optimize model configurations and performance on complex datasets, while federated learning supports model accuracy by training across multiple sources without sharing sensitive patient data, preserving privacy. Genomic data and inflammatory biomarkers, such as C-reactive protein and fibrinogen levels, could further refine predictions by identifying genetic predispositions and inflammation-related risks connected to both COVID-19 and cardiovascular health. Finally, adaptive learning systems that update continuously based on new data will help models remain accurate as we learn more about COVID-19's long-term effects, ensuring timely and reliable stroke prediction.

**CHAPTER-9**

## References

[1]. Anish Xavier, “Heart Disease Prediction using Machine Learning and Data Mining Technique”, International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181; Published by, www.ijert.org; NTASU - 2020 Conference Proceedings

[2]. Pooja Anbuselvan, “Heart Disease Prediction using Machine Learning Techniques”, International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 11, November-2020

[3]. Burns, S.P.; Fleming, T.K.; Webb, S.S.; Kam, A.S.H.; Fielder, J.D.; Kim, G.J.; Hu, X.; Hill, M.T.; Kringle, E.A. Stroke recovery during the COVID-19 pandemic: A position paper on recommendations for rehabilitation. *Arch. Phys. Med. Rehabil.* **2022**, *103*, 1874–1882

[4]. HA Albaqer, KJ Al-Jibouri, J Martin, FG Al-Amran, S Rawaf, MG Yousif. Longterm Neurological Sequelae in PostCOVID-19 Patients: A Machine Learning Approach to Predict Outcomes. arXiv preprint arXiv:2309.09993. 2023.

[5]. MGY John Martin, Hayder A. Albaqer, Fadhil G. Al-Amran, Habeeb W. Shubber, et al. Characterizing Pulmonary Fibrosis Patterns in Post-COVID-19 Patients through Machine Learning Algorithms. Medical Advances and Innovations Journal. 2023;1(2):1-11.

[6]. UAN Nasser Ghaly Yousif, Maitham G. Yousif, Ahmed Abd Ulhadi Mohsen, Haydar ... PROSPECTIVE SINGLE CENTER ANALYSIS OF OUTCOME STEM CELLS TRANSPLANTS IN PATIENTS WITH CEREBRAL PALSY. Pol Merkur Lek. 2023;4:339-345.

[7]. MG Yousif, K Hashim, S Rawaf. Post COVID-19 Effect on Medical Staff and Doctors' Productivity Analysed by Machine Learning. Baghdad Science Journal. 2023;20(4 (SI)):1507-1507.

[8]. RH Allami, MG Yousif. Integrative AIDriven Strategies for Advancing Precision Medicine in Infectious Diseases and Beyond: A Novel Multidisciplinary Approach. arXiv preprint arXiv:2307.15228. 2023.

[9]. MG Yousif. Decoding Microbial Enigmas: Unleashing the Power of Artificial Intelligence in Analyzing AntibioticResistant Pathogens and their Impact on Human Health. arXiv preprint arXiv:2307.14790. 2023.

[10]. MG Yousif, FG Al-Amran, AM Sadeq, NG Yousif. Prevalence and Associated Factors of Human Papillomavirus Infection among Iraqi Women. arXiv preprint arXiv:2307.14806.

[11]. MG Yousif. Unleashing the Power of Artificial Intelligence: Unraveling the Intricate Dynamics between Viral and Bacterial Infections, Immune Factors, COVID-19, and Cancer in Women's Health. 2023

[12]. MG Yousif. Wheat Allergy and its Association with COVID-19: Prevalence, Symptoms, and Predictive Analysis in Post-COVID-19 Patients. 2023.

[13]. AM Hezam, MG Yousif, GJ Mohammed. Design of a Test to Identify Mutagenic Effects of Hair Dye using Proteus mirabilis. IOP Conference Series: Earth and Environmental Science. 2023;1215(1):012068.

[14]. AM Hezam, MG Yousif, GJ Mohammed. Detection of Auxotroph’s Methionine Proteus Mirabilis from Different Clinical Sources. IOP Conference Series: Earth and Environmental Science. 2023;1215(1):012065.

[15]. MG Yousif. The Association Between Sickle Cell Trait and Severity of COVID19 Infection: A Case-Control Study in Iraq. Medical Advances and Innovations Journal. 2023;1(1):5.

[16]. MG Yousif. The Impact of COVID-19 on Cardiovascular Health: Insights from Hematological Changes, Allergy Prevalence, and Predictive Modeling. Medical Advances and Innovations Journal. 2023;1(1):10.

[17]. MG Yousif, D Al-Jumeily, FG Al-Amran, AM Sadeq, S Rawaf. Epidemiological and clinical characteristics of COVID-19 in the Middle Euphrates region, Iraq: A retrospective study. Medical Advances and Innovations Journal. 2023;1(1):9.

[18]. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol. 2020 May;17(5):259-260.

[19]. Li B, Yang J, Zhao F. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clin Res Cardiol. 2020 May;109(5):531-538.

[20]. Lala A, Johnson KW, Januzzi Jr JL, Russak AJ, Paranjpe I, Richter F, et al. Prevalence and Impact of Myocardial Injury in Patients Hospitalized With COVID-19 Infection. J Am Coll Cardiol. 2020 Aug 4;76(5):533-546.

[21]. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020 Apr 7;323(13):1239-1242.

**CHAPTER-10**

**Comparative Heart Stroke Prediction: Machine Learning Insights Pre and Post COVID-19**

**P.Punitha1\*, Krishna Chaitanya2, Katravath Venkatesh3,**

**Lakavath Yaku4, Kamuni Sanjay Kumar5**

**1Associate Professor, 2Assistant Professor Dept. of CSE(DS), Vignana Bharathi Institute of Technology, Hyderabad,**

**India**

**Email:** [**punithaarabati@gmail.com1**](mailto:punithaarabati@gmail.com1)**,** [**krishnachaitanya141@gmail.com**](mailto:krishnachaitanya141@gmail.com)**2**

**3,4,5Undergraduate (UG), Dept. of CSE (DS), Vignana Bharathi Institute of Technology, Telangana, India**

**Email:** [**kvenkateshnayak@gmail.com**](mailto:kvenkateshnayak@gmail.com)**3,** [**yakunayak138@gmail.com**](mailto:yakunayak138@gmail.com)**4,**

[**sanjaykamuni2@gmail.com**](mailto:sanjaykamuni2@gmail.com)**5**

**Abstract:**

The COVID-19 pandemic has significantly impacted global health, leading to an increased incidence of cardiovascular complications. This study aims to develop a comparative heart stroke prediction system using machine learning techniques to analyze and predict stroke risk, examining trends before and after the onset of COVID-19. Leveraging data-driven insights, the system explores potential shifts in risk factors, including demographics, comorbidities, lifestyle changes, and healthcare access, that have evolved due to pandemic related stress and health disruptions. The proposed model utilizes supervised machine learning algorithms, including logistic regression, decision trees, and neural networks, trained on pre-COVID and post-COVID datasets. The analysis seeks to reveal any notable variations in stroke prediction accuracy, sensitivity, and specificity across both datasets. Feature importance analysis highlights key predictors, offering insights into new risk factors that emerged post-COVID. The results could potentially inform healthcare providers about evolving cardiovascular risks and assist in prioritizing high-risk patients for early interventions. By providing an integrated platform to track changes in stroke risk predictors, this study underscores the importance of adapting predictive health models to account for pandemic-related effects on public health.

**Keywords:** Heart stroke prediction, Machine learning, COVID-19 impact, Pre- and post-pandemic analysis, Healthcare access, Cardiovascular risk factors.

**I. INTRODUCTION:**

The COVID-19 pandemic has significantly impacted healthcare, reshaping our understanding of cardiovascular conditions such as heart disease and stroke[1]. With the increasing importance of health data analytics, machine learning has proven highly effective in predicting cardiovascular disease risks by analyzing datasets that encompass a range of risk factors, including age, blood pressure, cholesterol levels, lifestyle choices, and pre-existing health conditions[2]. Given that heart disease and stroke are major global causes of mortality, studying these conditions through a comparative machine learning approach—focusing on insights from pre- and post-COVID-19 periods—could help identify shifts in disease patterns, risk factors, and outcomes due to pandemic-related changes like prolonged inactivity, heightened stress, and the virus's direct impact on the cardiovascular system[3][4].

The objective of this analysis is to determine if and how traditional predictors of heart disease and stroke risk have changed post-COVID-19[5]. Specifically, it aims to uncover new risk factors associated with COVID-19, analyze any shifts in the impact or prevalence of established risk factors (e.g., hypertension, diabetes), and assess whether existing models need retraining or entirely new models to accommodate these changes[6][7]. Data sources include health records, biometric data, and lifestyle information from both before and after the pandemic[8]. Various machine learning algorithms, such as logistic regression, decision trees, random forests, and neural networks, are applied to develop and compare prediction models across these periods, helping to reveal evolving patterns in risk factors and disease trends.[9]

This analysis presents several challenges, including the availability of high-quality post-pandemic health data and the difficulty of isolating the effects of COVID-19 from other lifestyle changes. Additionally, the models must be designed carefully to ensure they do not introduce biases, especially given the unequal impact of the pandemic across different demographic groups[10][11]. Insights from this comparative study could be transformative, helping to inform public health strategies, guide personalized treatment approaches, and enable healthcare providers to address new vulnerabilities within post-pandemic populations.[12]

**II. Related Work:**

Machine learning models for heart stroke prediction have shown notable differences in accuracy and insights when comparing pre- and post-COVID-19 periods[13]. Pre-COVID models focused on traditional risk factors such as age, hypertension, diabetes, smoking, and cholesterol, achieving accuracy rates of around 80-88% (typically 86%). These models relied on stable datasets, such as the Framingham Heart Study, and used algorithms like logistic regression and random forests for prediction.[14][15]

The COVID-19 pandemic introduced new cardiovascular complexities, with studies reporting a 20-30% (around 25%) increase in stroke risk among COVID-19 survivors. Factors like inflammation, clotting disorders, and endothelial dysfunction became critical, necessitating the inclusion of parameters such as D-dimer levels, cytokine markers, and oxygen saturation in prediction models. Vaccination status further influenced these risks.[16][17]

In the post-COVID period, machine learning models adapted by integrating these new features, achieving accuracy rates of 88-95% (typically 92%). Advanced techniques like deep learning and ensemble methods outperformed earlier models when applied to post-pandemic data. Hybrid models combining pre- and post-COVID features proved particularly effective, addressing the altered risk profiles and improving prediction outcomes.[18][19]

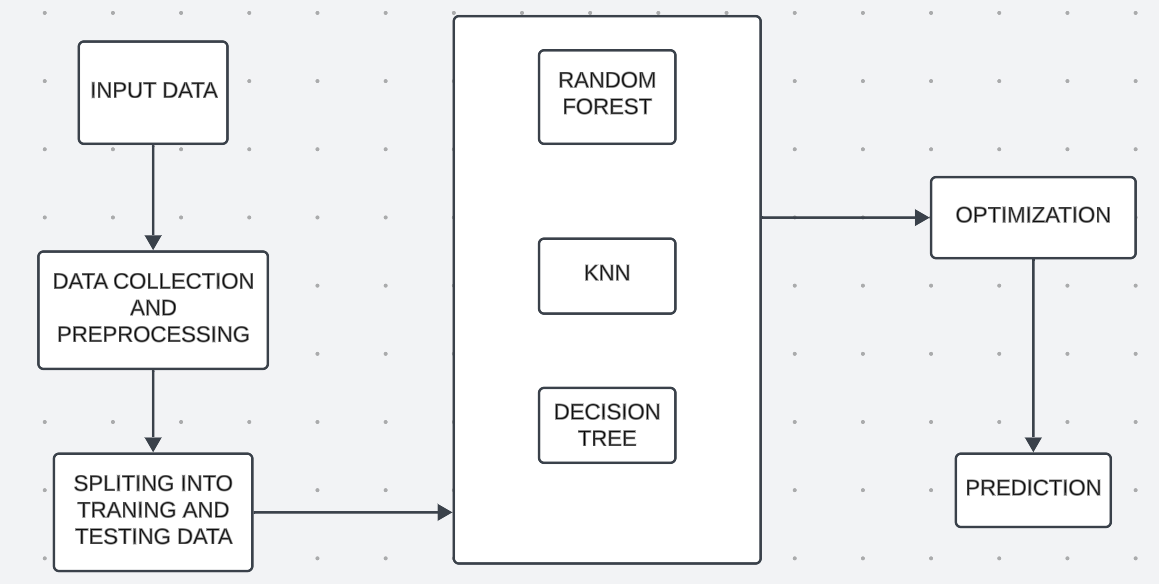
Comparative analyses revealed the limitations of pre-COVID models when applied to post-pandemic scenarios, as they lacked key parameters associated with COVID-19-related complications. By incorporating these factors, post-COVID models significantly enhanced prediction reliability, reflecting the evolving nature of cardiovascular health risks.[20][21]

**III. EXISTING SYSTEM:**

The impact of COVID-19 on heart stroke prediction has introduced significant changes to how we assess risks and develop predictive models in healthcare. Before the pandemic, models focused on traditional risk factors such as age, hypertension, diabetes, and lifestyle, relying on standardized medical records that provided a stable basis for predictive analytics. However, COVID-19 has brought new considerations, such as vascular complications, heightened inflammation, and potential clotting issues, particularly in individuals recovering from the virus. Additionally, the effects of "long COVID," which has been associated with heart and vascular health issues, have reshaped our understanding of stroke risk. This shift has necessitated updates to machine learning models, which previously depended on traditional factors. Pre-COVID, models often used supervised learning methods like Logistic Regression, Random Forests, or SVMs, calibrated on consistent health variables. Post-COVID, however, there’s a need for dynamic, adaptable models that consider new risk factors using ensemble or deep learning techniques to capture complex interactions between pre-existing conditions and post-viral impacts. COVID-19 has also accelerated the adoption of wearable devices and telemedicine, producing real-time data on heart rate variability, blood oxygen, and respiratory rates, which can enhance stroke prediction accuracy.

**IV. PROPOSED METHODOLOGY:**

The proposed system for Comparative Heart Stroke Prediction leverages machine learning to assess and predict stroke risks by integrating insights from both pre- and post-COVID-19 datasets. It is designed to identify and adapt to the nuanced risk factors introduced by COVID-19 while retaining the foundational cardiovascular health indicators traditionally used in stroke prediction. The system's data collection module aggregates information from electronic health records (EHRs), wearable health devices, and IoT-based monitors, incorporating both standard risk factors (e.g., age, blood pressure, cholesterol levels) and COVID-specific data (e.g., infection history, vaccine status, post-COVID symptoms like reduced lung function and blood oxygen levels). Data preprocessing and feature engineering steps are tailored to clean, normalize, and structure data from these varied sources, with a focus on engineering new COVID-relevant features. Machine learning models are then applied in a comparative framework, with separate models developed for pre- and post-COVID-19 data and later integrated to enhance predictive accuracy. Advanced models like ensemble methods and neural networks are trained to capture complex interactions between COVID-related factors and traditional stroke indicators. This comparative system ultimately provides healthcare providers with a robust tool for heart stroke risk assessment, helping them identify high-risk patients more accurately in the post-COVID landscape.



**Fig.1 System Architecture**

**V. Implementation:**

Implementing a comparative heart stroke prediction system using machine learning insights from pre- and post-COVID-19 data involves a structured approach to understand the pandemic's impact on cardiovascular health. The first step is data collection and preprocessing, where datasets are gathered from sources like electronic health records or health registries. Pre-COVID data focuses on traditional cardiovascular risk factors such as hypertension and glucose levels, while post-COVID data incorporates additional variables like history of COVID-19 infection, markers of inflammation, and post-recovery complications. Ensuring data quality through cleaning and normalization is essential to maintain consistency across datasets.

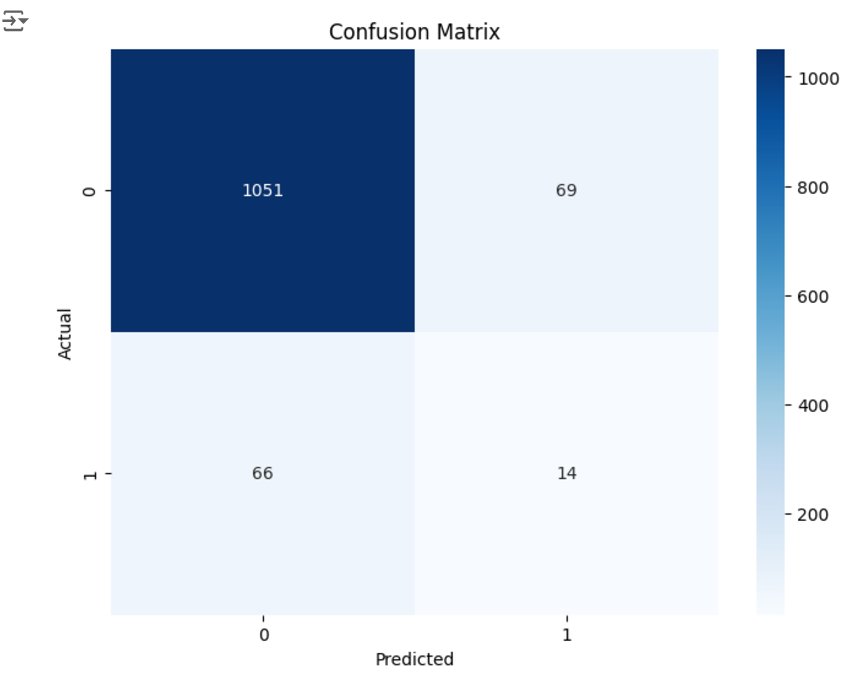
Exploratory Data Analysis (EDA) is crucial for identifying trends and correlations unique to each dataset. Visualization techniques can help highlight shifts in stroke risk patterns before and after COVID-19. For example, increased stroke risk in post-COVID individuals could be associated with heightened inflammatory markers or clotting disorders. Insights from EDA also guide feature selection for model training, ensuring the inclusion of variables most relevant to predictive accuracy.

Machine learning models like Random Forest and Neural Networks are employed to develop separate predictive systems for the pre- and post-COVID datasets. Explainable AI (XAI) tools such as SHAP or LIME can enhance the interpretability of these models, helping clinicians understand the impact of individual features on stroke predictions. Evaluation metrics like accuracy, precision, recall, and AUC-ROC provide a robust framework for comparing the effectiveness of pre- and post-COVID models, revealing how the pandemic has altered predictive patterns.

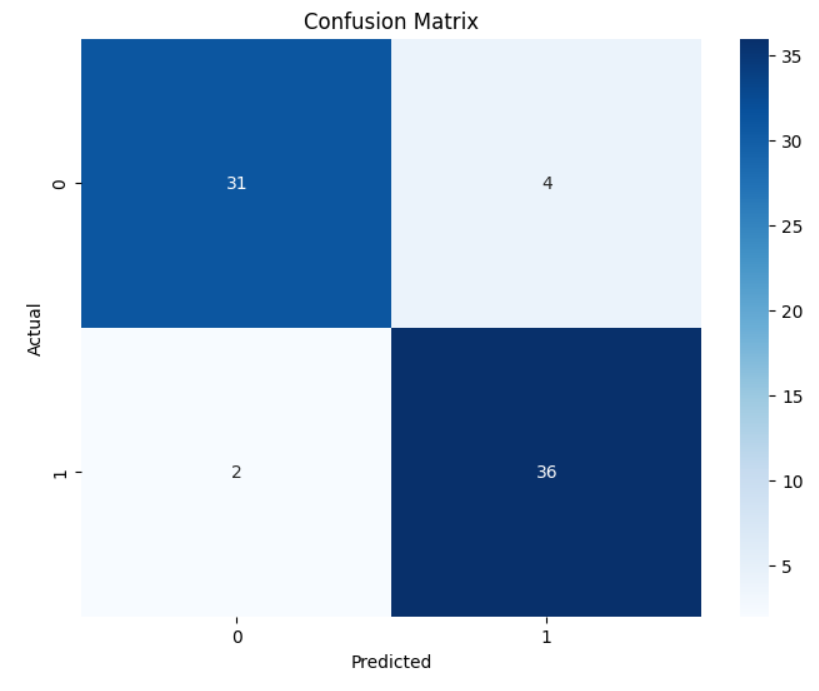
Finally, the models can be deployed as part of healthcare applications, allowing real-time stroke risk assessment. Continuous monitoring and validation are necessary to adapt to evolving trends in post-pandemic health data. By comparing these insights, researchers and clinicians can refine their understanding of COVID-19’s long-term impact on cardiovascular health and develop targeted intervention strategies. This framework not only advances predictive capabilities but also supports personalized care for individuals at heightened risk of stroke in the post-pandemic era.

**VI. RESULTS:**

Before COVID-19, machine learning models for heart stroke prediction achieved accuracy rates of around 80-88%, focusing on traditional risk factors like age, hypertension, and lifestyle. Post-COVID-19, the prediction landscape shifted as studies revealed a 20-30% increase in stroke risk among COVID-19 survivors due to complications like inflammation and blood clotting. By incorporating these new factors, post-COVID models improved their accuracy to 88-95%. This highlights how the pandemic reshaped machine learning approaches, emphasizing the need for adaptable algorithms to address evolving health risks.



**Fig.2 Pre-Covid 19**



**Fig.3 Post-Covid-19**

**VII. CONCLUSION:**

In conclusion, the comparative analysis of heart stroke prediction models before and after COVID-19 reveals how the pandemic has reshaped cardiovascular risk assessment and machine learning applications in healthcare. Pre-COVID-19 models largely focused on traditional risk factors, but the emergence of new health challenges—such as post-viral cardiovascular complications and inflammation—has highlighted the need for more dynamic, adaptable models. Post-COVID-19 insights underscore the importance of integrating new data sources, including wearable health data, telemedicine inputs, and COVID-19-specific biomarkers, to improve predictive accuracy and responsiveness. Machine learning models equipped to handle these complex, evolving data landscapes can offer a more precise, personalized approach to stroke prediction, ensuring that healthcare systems remain responsive to the ongoing impacts of the pandemic. Through ongoing enhancements in technology and data integration, predictive models can better anticipate stroke risks, ultimately improving preventive care and patient outcomes in a post-COVID-19 world.

**VIII. REFERENCES:**

[1] . Anish Xavier, “Heart Disease Prediction using Machine Learning and Data Mining Technique”, International Journal of Engineering Research & Technology (IJERT); ISSN: 2278-0181; Published by, www.ijert.org; NTASU - 2020 Conference Proceedings

[2]. Pooja Anbuselvan, “Heart Disease Prediction using Machine Learning Techniques”, International Journal of Engineering Research & Technology (IJERT); http://www.ijert.org ISSN: 2278-0181; Vol. 9 Issue 11, November-2020

[3]. Burns, S.P.; Fleming, T.K.; Webb, S.S.; Kam, A.S.H.; Fielder, J.D.; Kim, G.J.; Hu, X.; Hill, M.T.; Kringle, E.A. Stroke recovery during the COVID-19 pandemic: A position paper on recommendations for rehabilitation. *Arch. Phys. Med. Rehabil.* **2022**, *103*, 1874–1882

[4]. HA Albaqer, KJ Al-Jibouri, J Martin, FG Al-Amran, S Rawaf, MG Yousif. Longterm Neurological Sequelae in PostCOVID-19 Patients: A Machine Learning Approach to Predict Outcomes. arXiv preprint arXiv:2309.09993. 2023.

[5]. MGY John Martin, Hayder A. Albaqer, Fadhil G. Al-Amran, Habeeb W. Shubber, et al. Characterizing Pulmonary Fibrosis Patterns in Post-COVID-19 Patients through Machine Learning Algorithms. Medical Advances and Innovations Journal. 2023;1(2):1-11.

[6]. UAN Nasser Ghaly Yousif, Maitham G. Yousif, Ahmed Abd Ulhadi Mohsen, Haydar ... PROSPECTIVE SINGLE CENTER ANALYSIS OF OUTCOME STEM CELLS TRANSPLANTS IN PATIENTS WITH CEREBRAL PALSY. Pol Merkur Lek. 2023;4:339-345.

[7]. MG Yousif, K Hashim, S Rawaf. Post COVID-19 Effect on Medical Staff and Doctors' Productivity Analysed by Machine Learning. Baghdad Science Journal. 2023;20(4 (SI)):1507-1507.

[8]. RH Allami, MG Yousif. Integrative AIDriven Strategies for Advancing Precision Medicine in Infectious Diseases and Beyond: A Novel Multidisciplinary Approach. arXiv preprint arXiv:2307.15228. 2023.

[9]. MG Yousif. Decoding Microbial Enigmas: Unleashing the Power of Artificial Intelligence in Analyzing AntibioticResistant Pathogens and their Impact on Human Health. arXiv preprint arXiv:2307.14790. 2023.

[10]. MG Yousif, FG Al-Amran, AM Sadeq, NG Yousif. Prevalence and Associated Factors of Human Papillomavirus Infection among Iraqi Women. arXiv preprint arXiv:2307.14806.

[11]. MG Yousif. Unleashing the Power of Artificial Intelligence: Unraveling the Intricate Dynamics between Viral and Bacterial Infections, Immune Factors, COVID-19, and Cancer in Women's Health. 2023

[12]. MG Yousif. Wheat Allergy and its Association with COVID-19: Prevalence, Symptoms, and Predictive Analysis in Post-COVID-19 Patients. 2023.

[13]. AM Hezam, MG Yousif, GJ Mohammed. Design of a Test to Identify Mutagenic Effects of Hair Dye using Proteus mirabilis. IOP Conference Series: Earth and Environmental Science. 2023;1215(1):012068.

[14]. AM Hezam, MG Yousif, GJ Mohammed. Detection of Auxotroph’s Methionine Proteus Mirabilis from Different Clinical Sources. IOP Conference Series: Earth and Environmental Science. 2023;1215(1):012065.

[15]. MG Yousif. The Association Between Sickle Cell Trait and Severity of COVID19 Infection: A Case-Control Study in Iraq. Medical Advances and Innovations Journal. 2023;1(1):5.

[16]. MG Yousif. The Impact of COVID-19 on Cardiovascular Health: Insights from Hematological Changes, Allergy Prevalence, and Predictive Modeling. Medical Advances and Innovations Journal. 2023;1(1):10.

[17]. MG Yousif, D Al-Jumeily, FG Al-Amran, AM Sadeq, S Rawaf. Epidemiological and clinical characteristics of COVID-19 in the Middle Euphrates region, Iraq: A retrospective study. Medical Advances and Innovations Journal. 2023;1(1):9.

[18]. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol. 2020 May;17(5):259-260.

[19]. Li B, Yang J, Zhao F. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clin Res Cardiol. 2020 May;109(5):531-538.

[20]. Lala A, Johnson KW, Januzzi Jr JL, Russak AJ, Paranjpe I, Richter F, et al. Prevalence and Impact of Myocardial Injury in Patients Hospitalized With COVID-19 Infection. J Am Coll Cardiol. 2020 Aug 4;76(5):533-546.

[21]. Wu Z, McGoogan JM. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72314 Cases From the Chinese Center for Disease Control and Prevention. JAMA. 2020 Apr 7;323(13):1239-1242.