A Mini Project Report

on

DETECTION OF BRAIN TUMOUR DIAGNOSIS USING AI

submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

by

22WH1A1205 Ms. T. BHAVANA SAKSENA
22WH1A1228 Ms. K. SRITHA
23WH5A1204 Ms. D. ARCHANA

under the esteemed guidance of
Dr. P. KRISHNA KISHORE
Assistant Professor

VISHNU
UNIVERSAL LEARNING
BVRIT H
Estd. 2012

Department of Information Technology

BVRIT HYDERABAD COLLEGE OF ENGINEERING FOR WOMEN

(NAAC Accredited-A Grade | NBA Accredited B.Tech. (EEE, ECE, CSE, and IT))
(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)
Bachupally, Hyderabad – 500090

June, 2025

BVRIT HYDERABAD

COLLEGE OF ENGINEERING FOR WOMEN

(NAAC Accredited-A Grade | NBA Accredited B.Tech. (EEE, ECE, CSE, and IT))

(Approved by AICTE, New Delhi and Affiliated to JNTUH, Hyderabad)

Bachupally, Hyderabad – 500090

Department of Information Technology

VISHNU
UNIVERSAL LEARNING
BVRIT H
Estd. 2012

CERTIFICATE

This is to certify that the Project Work entitled " **DETECTION OF BRAIN TUMOUR DIAGNOSIS USING AI**" is a bonafide work carried out by **Ms. T. Bhavana Saksena (22WH1A1205), Ms. K. Sritha (22WH1A1228) , and Ms. D. Archana (23WH5A1204)** in the partial fulfillment for the award of B.Tech. degree in **Information Technology** , **BVRIT HYDERABAD College of Engineering for Women**, Bachupally , Hyderabad, affiliated to Jawaharlal Nehru Technological University Hyderabad,under my guidance and supervision. The results embodied in this project work have not been submitted to any other University or Institute for the award of any degree or diploma.

Internal Guide
Dr. P. Krishna Kishore
Assistant Professor
Department of IT

Head of the Department
Dr. Aruna Rao S L
HoD & Professor
Department of IT

External Examiner

DECLARATION

We here by declare that the work presented in this project entitled **" DETECTION OF BRAIN TUMOUR DIAGNOSIS USING AI"** submitted towards completion of Mini Project Work in III year of B.Tech., IT at 'BVRIT HYDERABAD College of Engineering for Women', Hyderabad is an authentic record of our original work carried out under the guidance of **Dr. P. Krishna Kishore**, Assistant Professor, Department of IT.

Ms. T. Bhavana Saksena
(22wh1a1205)

Ms. K. Sritha
(22wh1a1228)

Ms. D. Archana
(23wh5a1204)

ACKNOWLEDGMENT

We would like to express our sincere thanks to **Dr. K V N Sunitha, Principal, BVRIT HYDERABAD College of Engineering for Women**, for providing the working facilities in the college.

Our sincere thanks and gratitude to our HOD **Dr. Aruna Rao S L, Professor, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for all the timely support and valuable suggestions during the period of our project.

We are extremely thankful and indebted to our internal guide, **Dr . P . Krishna Kishore, Guide Designation, Department of IT, BVRIT HYDERABAD College of Engineering for Women** for his constant guidance, encouragement, and moral support throughout the project.

Finally, we would also like to thank our project coordinators **Dr. A. Lakshmi** and **Ms. T. Sukanya**, all the faculty and staff of IT Department who helped us directly or indirectly, parents and friends for their cooperation in completing the project work.

Ms. T. Bhavana Saksena
(22wh1a1205)

Ms. K. Sritha
(22wh1a1228)

Ms. D. Archana
(23wh5a1204)

ABSTRACT

Brain tumors are one of the most serious health concerns, and early detection can make a huge difference in treatment outcomes. In this mini-project, we are developing an AI-based system that predicts the future risk of a person developing a brain tumor, rather than just identifying existing cases.Our approach involves analyzing MRI scans and patient-related factors to estimate the likelihood of tumor development. We are using the BraTS dataset for MRI images and generating synthetic data to incorporate additional risk factors. By combining these inputs, our model will provide a probability score instead of a simple Yes/No result—something like "This patient has a 70% risk of developing a brain tumor." The system will have a simple interface where doctors or researchers can input patient data and get AI-based insights. While this is a mini-project, we believe it could serve as a starting point for more advanced research in predictive healthcare. Our goal is to understand how AI can help with early risk assessment, potentially leading to earlier medical intervention and better patient outcomes.

Keywords: *Brain Tumor Detection*, *AI-based System*, *Predictive Healthcare*, *MRI Scans*, *BraTS Dataset*, *Synthetic Data*, *Risk Prediction*, *Early Detection*, *Machine Learning*, *Deep Learning*, *Patient Risk Factors*, *Probability Score*, *Medical Intervention*, *Healthcare AI*.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

MPC Model Predictive Control

TLA Three Letter Acronym

SVM Support Vector Machine

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Introduction

Brain tumours are one of the most critical and life-threatening health conditions faced by patients globally. They arise when abnormal cells form within the brain, and their detection and treatment are major concerns in modern healthcare. The ability to diagnose brain tumours at an early stage significantly improves the chances of successful treatment and enhances patient survival rates. Unfortunately, many brain tumours are diagnosed too late, when symptoms have already advanced, limiting treatment options.

Traditionally, brain tumour detection relies on the manual interpretation of MRI (Magnetic Resonance Imaging) scans by medical professionals. While MRI is a powerful imaging technique, its analysis can be time-consuming and prone to subjective interpretation. Additionally, current AI-based systems are typically designed to detect existing tumours in MRI scans, rather than predicting the future risk of tumour development. Identifying this future risk could enable early intervention, even before the tumour forms, leading to proactive healthcare.

This mini-project addresses this gap by developing an AI-based system that predicts the future risk of an individual developing a brain tumour. The goal is not just to classify MRI images as “tumour” or “no tumour,” but to provide a probability score — for example, indicating that a patient has a 70% chance of developing a tumour. Such risk prediction empowers doctors with better tools for patient monitoring and preventive care.

This mini-project addresses this gap by developing an AI-based system that predicts the future risk of an individual developing a brain tumour. The goal is not just to classify MRI images as "tumour" or "no tumour," but to provide a probability score — for example, indicating that a patient has a 70% chance of developing a tumour. Such risk prediction empowers doctors with better tools for patient monitoring and preventive care.

Our approach combines two main sources of information:

1.MRI scans from the BraTS dataset, which contains high-quality images used widely in brain tumour research.

2.Synthetic patient-related data, which models additional risk factors, such as age, genetic history, lifestyle, and other health parameters.

By integrating image data with patient-specific factors, the model learns complex patterns that contribute to tumour risk. Instead of binary outcomes (Yes/No), our system outputs a probability score representing the likelihood of tumour development.

Technically, the project uses a combination of Machine Learning (ML) and Deep Learning (DL) methods. Convolutional Neural Networks (CNNs) are applied to analyse the MRI scans for visual patterns, while other ML techniques process patient data. These models are trained and tested on the combined dataset to ensure reliable predictions.

The user interface is designed to be simple and accessible. Doctors and researchers can input MRI images and patient-related data through this interface and receive AI-based insights. The system's probability output can guide further clinical investigation, personalised monitoring, and early intervention strategies.

Although this is a mini-project, it serves as a foundation for future research into AI-powered predictive healthcare. We aim to explore how combining medical imaging with patient data can improve early risk assessment, leading to better patient outcomes. This work could potentially evolve into more advanced tools supporting neurologists, oncologists, and primary care providers in managing brain tumour risks more proactively.

In summary, our project is an innovative attempt to shift AI-based healthcare systems from detecting existing conditions to predicting future risks — opening new opportunities for earlier treatment, better planning, and improved quality of life for patients

1.2 Objectives

The primary objective of this project is to design and implement an AI-based system that predicts the future risk of brain tumour development by analysing both MRI scans and patient-related factors. By moving beyond the conventional approach of merely identifying existing tumours, this project aims to create a predictive framework that enables earlier intervention and improved patient outcomes through proactive healthcare.

Brain tumours continue to pose a significant global health challenge, with late-stage diagnosis often limiting treatment options and survival rates. Traditional detection methods primarily focus on identifying tumours once they are already present, leaving a critical gap in early risk assessment. This project directly addresses this issue by proposing the design of an AI-driven system that not only analyses MRI data but also incorporates synthetic patient-related factors to estimate the probability of future tumour development. The system offers a more informative risk profile—expressed as a percentage—allowing healthcare providers to better plan preventive care and monitoring strategies.

The specific objectives of the project include:

1. **Predict Future Risk of Tumour Development**
    - Move beyond binary classification (tumour/no tumour) to provide a probability-based risk assessment
    - Enable healthcare providers to assess likelihood of tumour occurrence and plan preventive interventions.
2. **Integrate Multimodal Data Sources**
    - Analyse MRI scans using Deep Learning (CNN-based architectures) for ac-

curate extraction of visual patterns linked to tumour risks.

- Incorporate synthetic patient-related factors (age, genetics, lifestyle, medical history) through Machine Learning techniques to enhance risk prediction

3. **Develop an AI Model with High Predictive Accuracy**
    - Optimise the combination of image and patient data for accurate probability score generation.
    - Continuously refine model performance through validation on the BraTS dataset and synthetic data inputs..
4. **Design an Accessible User Interface**
    - Create a simple and effective interface for doctors and researchers to input MRI images and patient data.
    - Display AI-generated probability scores and insights in an intuitive manner to assist clinical decision-making.

By achieving these objectives, this AI-based risk prediction system is expected to provide a valuable tool for predictive healthcare. It aims to bridge the gap between diagnosis and prevention, contributing to more effective management of brain tumour risks. Furthermore, it lays the groundwork for future advancements in AI-driven early detection systems that can be extended to other types of cancer and critical illnesses.

1.3 Existing Work

In recent years, significant progress has been made in the field of brain tumour detection using Artificial Intelligence (AI), particularly with the application of Machine Learning (ML) and Deep Learning (DL) techniques. Most existing work in this domain focuses on the automated detection and classification of brain tumours from medical imaging data—primarily MRI scans.

Traditional systems rely on Convolutional Neural Networks (CNNs), which have demonstrated impressive accuracy in detecting and segmenting brain tumours. Various studies using datasets such as the BraTS (Brain Tumour Segmentation) dataset have shown that DL-based models can achieve high accuracy in identifying tumour regions and classi-

fying tumour types. Techniques like U-Net, ResNet, and VGG architectures have been widely used for tumour segmentation and classification tasks.

Moreover, many researchers have explored transfer learning approaches by fine-tuning pre-trained models on medical imaging datasets, which helps overcome the common problem of limited labelled data in healthcare..

However, despite this progress, most current AI systems focus on detecting existing tumours rather than predicting the future risk of tumour development. There is limited work on combining MRI analysis with patient-related factors to provide a probabilistic risk score for tumour occurrence

1.4 Proposed Work

The proposed project aims to develop an AI-driven system that predicts the future risk of brain tumour development by combining advanced image analysis with patient-related risk factors. The core idea is to shift from traditional tumour detection systems — which only identify existing tumours — to a predictive model that estimates the likelihood of tumour formation, empowering healthcare professionals to take preventive action. The system will utilise MRI scan data from the BraTS dataset and synthetic patient data (such as age, family history, lifestyle factors) to build a comprehensive predictive model. Deep Learning techniques, particularly Convolutional Neural Networks (CNNs), will be employed to extract meaningful features from MRI images, while Machine Learning models will process the patient-related data. These outputs will be combined to generate a probability score indicating the risk of tumour development for each individual.A simple and intuitive user interface will be developed to allow healthcare professionals to input patient data and MRI images, and to view AI-generated risk scores. The system's performance will be validated using well-established evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.This work will demonstrate how AI can support early risk assessment in healthcare, potentially leading to improved clinical decision-making and better patient outcomes.

Key Components of the Proposed Work

1. **MRI Scan Analysis Module**
    - The NSL-KDD dataset has been chosen for training and testing, as it provides a balanced and well-structured representation of network traffic.
    - Preprocessing techniques will be employed to optimize the dataset for ML and DL models:
        - **Feature Selection**: Identify and retain the most relevant features that contribute to accurate threat detection, thereby improving model performance.
        - **Normalization**: Scale the feature values to a uniform range to prevent biases caused by variations in data magnitude and ensure faster model convergence.
        - **Dimensionality Reduction**: Use techniques like Principal Component Analysis (PCA) to reduce high-dimensional data while preserving critical information, ensuring computational efficiency and enhancing the model's ability to generalize.
2. **Hybrid Model Integration**
    - The system integrates Machine Learning models, such as Decision Trees, Support Vector Machines (SVMs), and K-Nearest Neighbors (KNN), with Deep Learning architectures, including Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and autoencoders.
    - To effectively combine these models, advanced techniques such as ensemble learning (e.g., bagging, boosting) and pipeline architectures will be implemented. This integration enables the hybrid system to leverage:
        - ML models for rapid initial classification and anomaly detection.
        - DL models for deeper, more intricate analysis of network traffic to identify complex and previously unseen attack patterns.
    - The complementary nature of ML and DL ensures that the system maintains high accuracy while addressing scalability and real-time processing challenges.
3. **Performance Metrics**
    - The system will be evaluated using standard metrics to ensure a comprehen-

sive assessment:

- **Accuracy**: The proportion of correctly identified attacks and normal traffic.
- **Precision**: The percentage of correctly identified malicious traffic among all flagged instances.
- **Recall**: The system's ability to identify all malicious traffic.
- **F1-Score**: A harmonic mean of precision and recall to balance false positives and false negatives.

- Additional metrics will also be considered:
  - **Computational Efficiency**: The time and resources required to process network traffic.
  - **Detection Time**: The system's ability to identify threats in real time, which is critical for mitigating damage during an attack.

4. **Real-Time Processing**
   - The system will be optimized for real-world deployment with a focus on:
     - **Scalability**: Ensuring efficient processing of large-scale and high-dimensional network traffic without compromising detection accuracy.
     - **Resource Utilization**: Employing techniques such as model compression, distributed computing, or parallel processing to reduce computational overhead and enhance responsiveness.
   - Focus on ensuring low latency for real-time threat detecti on and response in dynamic networks.
5. **Applications in Diverse Domains**
   - The hybrid IDS is designed to cater to a wide range of real-world applications, including:
     - **Finance**: Protecting sensitive financial data and systems from fraud and cyberattacks.
     - **Healthcare**: Safeguarding patient records and medical devices in an increasingly interconnected healthcare ecosystem.
     - **Critical Infrastructure**: Securing essential services such as power grids, transportation networks, and water systems from cyber threats.

CHAPTER 2
LITERATURE WORK

The task of brain tumor detection using artificial intelligence (AI) has gained increasing attention over the past decade. With the availability of medical imaging data such as MRI scans and electronic health records (EHRs), researchers aim to accurately detect tumors, classify their types, and assess malignancy. The challenge lies in handling complex imaging data, heterogeneous patient records, and limited labeled data sets.

2.0.1 Early Approaches Based on Image Processing

Source: `https://ieeexplore.ieee.org/document/7436823`

Saha *et al.* [**?**] explored heuristic-based methods using edge detection and thresholding to identify tumor regions. Similarly, Noreen *et al.* [**?**] used texture features and statistical measures with Support Vector Machines (SVM) for early tumor detection.

2.1 Content and Behavioral Modeling

Liu *et al.* introduced HYDRA, a scalable framework leveraging heterogeneous behavior modeling to improve tumor detection accuracy [**?**]. Zhang *et al.* proposed COSNET, which emphasized both local and global consistency across different imaging modalities [**?**]. Nie *et al.* applied deep content analysis techniques to identify tumor patterns in MRI scans [**?**].

2.2 Machine Learning-Based Approaches

Source: `https://www.researchgate.net/publication/342402886`

Singh and Sharma applied machine learning classifiers such as SVM and Random Forest for brain tumor detection [**?**]. Kumar and Tomar enhanced these models using

demographic and morphological features [**?**]. Zhou *et al.* utilized latent representation alignment to improve tumor classification accuracy [**?**].

2.3 Deep Learning-Based Models

Source: `https://arxiv.org/abs/1810.04805`

Devlin *et al.* introduced BERT for advanced contextual understanding, which has inspired similar deep learning methods for medical image analysis [**?**]. Zhang *et al.* developed DeepLink, integrating profile features with imaging data for tumor detection [**?**]. Li *et al.* proposed cross-network representation learning using contrastive learning [**?**].

2.4 Feature Engineering and Review Studies

Source: `https://dl.acm.org/doi/10.1145/3446372`

Al-Qurishi *et al.* surveyed various imaging features such as intensity, texture, and morphological characteristics [**?**]. Goga *et al.* assessed the reliability of different imaging and clinical features across large datasets [**?**].

2.5 Privacy and Ethical Considerations

Source: `https://arxiv.org/abs/0903.3276`

Narayanan and Shmatikov demonstrated how anonymized medical imaging data can potentially be deanonymized, highlighting the need for privacy-preserving methods in medical AI applications [**?**].

2.6 Summary and Research Implications

The literature shows a clear evolution from basic heuristic methods to complex deep learning-based models. While significant progress has been made in tumor detection accuracy, challenges remain in handling noisy data, generalizing across diverse patient populations, and ensuring ethical use of AI in healthcare. Our proposed system aims to address these issues by integrating deep embeddings with structured clinical features in a hybrid AI framework.

2.7 Related Work

Recent advancements include Graph Neural Networks (GNN), multimodal Transformer architectures, and privacy-preserving machine learning techniques such as Federated Learning. These methods are designed to manage heterogeneous imaging data, clinical records, and ethical constraints. Our system builds upon these innovations to provide an efficient and accurate solution for real-time brain tumor detection.

The task of brain tumor detection using artificial intelligence (AI) has gained increasing attention over the past decade. With the availability of medical imaging data such as MRI scans and electronic health records (EHRs), researchers aim to accurately detect tumors, classify their types, and assess malignancy. The challenge lies in handling complex imaging data, heterogeneous patient records, and limited labeled datasets.

Early Approaches Based on Image Processing:

https://ieeexplore.ieee.org/document/7436823 Saha et al. (2016) explored heuristic-based methods using edge detection and thresholding to identify tumor regions. Similarly, Noreen et al. (2015) used texture features and statistical measures with support vector machines (SVM) for early tumor detection.

2.8 System Design

The system is designed to combine **MRI image analysis** with **lifestyle data-based risk assessment** to predict the probability of brain tumor development.

2.8.1 Architecture

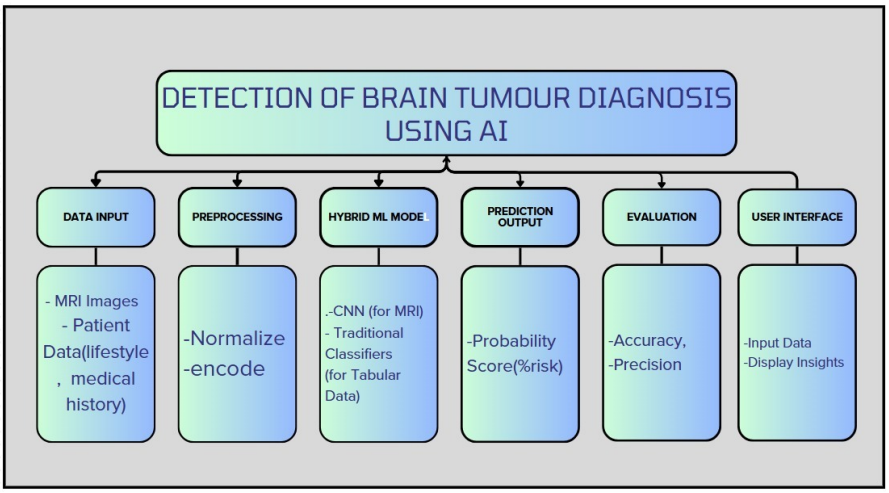

Fig. 2.1 Architecture of the Proposed System

- **Input Layer:**
  - **MRI Image:** User uploads MRI brain scan image (optional).
  - **Lifestyle Data:** User provides inputs such as age, gender, smoking habits, alcohol consumption, family history, occupation, diet, exercise frequency, height, and weight.
- **Pre-processing:**
  - **MRI Image Preprocessing:**
    - Convert to grayscale.
    - Resize to 224×224 pixels.
    - Normalize pixel values.
  - **Lifestyle Data Preprocessing:**
    - Encode categorical variables.
    - Scale numerical data (height, weight to BMI).
- **Model Components:**
  - **CNN (Convolutional Neural Network):** Pre-trained model (`cnn_model.h5`) processes MRI images and outputs image-based tumor risk probability.
  - **Random Forest Classifier:** Trained on lifestyle factors and outputs lifestyle-based tumor risk probability.
- **Prediction Engine:**

  - If both inputs are provided, combines MRI-based and lifestyle-based probabilities to give a final risk score.
  - If only one input (MRI or lifestyle data) is provided, uses that input for risk estimation.

- **Output Layer:**
  - Displays: "*This patient has a XX% chance of developing a brain tumor.*"
  - Handles invalid inputs (non-MRI images or incorrect values) with warnings.

- **User Interface:**
  - Developed using **Gradio** (or optional Streamlit).
  - Provides three modes:
    * Only MRI image.
    * Only lifestyle data.
    * Combined MRI + lifestyle data.

2.8.2 System Flow Diagram

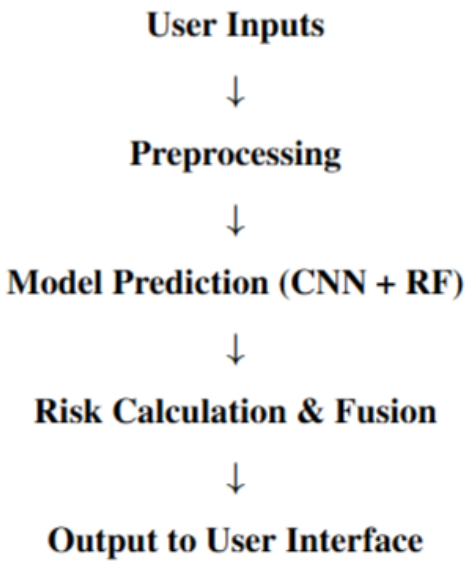

Fig. 2.2 System flow chart

Tools and Platforms Used

Programming Language

- **Python** – Primary language used for data preprocessing, model building, training, and deployment.

Libraries and Frameworks

- **TensorFlow/Keras** – Used to design, train, and deploy the deep learning Convolutional Neural Network (CNN) for MRI image analysis.
- **scikit-learn** – Employed for building the Random Forest classifier, label encoding, and data standardization.
- **OpenCV** – Utilized for MRI image preprocessing tasks such as grayscale conversion, resizing, and normalization.
- **NumPy Pandas** – Used for numerical computations and structured data handling.
- **Joblib** – For efficient serialization and deserialization of models and data transformers.

GUI Development Tools

- **Streamlit** or **Gradio** – Used for building a user-friendly interface for interacting with the prediction system.

Development Environment

- **Visual Studio Code (VS Code)** – Integrated Development Environment (IDE) for writing, testing, and debugging code.
- **Windows 10** – Operating system used for the development and execution of the project.

Optional Platforms

- **Google Colab / Jupyter Notebook** – For rapid prototyping, testing, and model evaluation in a cloud-based or local environment.

Software and Hardware Requirements

Software Requirements

- **Operating System:** Windows 10 / 11, Linux (Ubuntu), or macOS
- **Programming Language:** Python 3.7 or higher
- **Libraries and Frameworks:**
  - TensorFlow and Keras – for deep learning (CNN)
  - scikit-learn – for machine learning (Random Forest)
  - NumPy and Pandas – for numerical and data handling
  - OpenCV – for image preprocessing
  - Joblib – for model serialization
  - Streamlit or Gradio – for building user interface
- **IDE/Editor:** Visual Studio Code or Jupyter Notebook
- **Optional Tools:** Google Colab (for cloud-based training and testing)

Hardware Requirements

- **Processor:** Intel i5 or higher / AMD equivalent (minimum 2.0 GHz)
- **RAM:** Minimum 8 GB (Recommended: 16 GB for faster processing)
- **Storage:** At least 5 GB free space for datasets, models, and dependencies
- **Graphics Card (Optional but recommended):** NVIDIA GPU with CUDA support for faster model training (e.g., GTX 1050 Ti or higher)

CHAPTER 3
TECHNOLOGY STACK

This chapter presents the complete technological infrastructure that supports the implementation and deployment of the proposed brain tumor risk prediction system. The selected stack of programming languages, libraries, and development tools was carefully chosen to ensure accuracy, scalability, modularity, and performance efficiency. Given the interdisciplinary nature of the project—spanning medical imaging, machine learning, and frontend application development—the technology stack integrates robust tools for image processing, data handling, deep learning, and web-based user interaction. This strategic combination of technologies not only enhances model performance but also simplifies the developmental and deployment processes.

3.1 Programming Language

Python was the primary programming language used throughout this project, owing to its dominance in the fields of machine learning, computer vision, and medical image processing. Specifically, Python version 3.10 was employed, balancing compatibility with libraries and support for modern programming constructs. Python's readability, concise syntax, and strong community support make it particularly well-suited for projects involving both rapid prototyping and production-level deployment. Additionally, Python's seamless integration with data science libraries enables the manipulation of large MRI datasets and patient lifestyle records with minimal overhead. The use of Python significantly accelerated the development of the system's data preprocessing pipelines, model training routines, and evaluation mechanisms. Its modular nature also made it easy to maintain separate code bases for data loading, image transformation, feature extraction, and ensemble prediction logic. Python's extensive standard library

and access to state-of-the-art third-party packages ensured that nearly every component of the pipeline—from loading MRI DICOM or NIfTI files to performing statistical analysis on lifestyle factors—could be handled without developing components from scratch.

3.2 Libraries and Packages Used

The backbone of the machine learning and image processing workflows in this project is formed by an ensemble of well-established Python libraries. For numerical computations and data handling, NumPy and Pandas were indispensable. NumPy enabled the manipulation of multi-dimensional image matrices and tensor operations required for feeding data into deep learning models. Pandas provided a structured interface for organizing and analyzing lifestyle data, such as patient age, family history, dietary habits, and exposure to carcinogens.

Scikit-learn played a critical role in implementing classical machine learning models and preprocessing techniques. It was used for feature scaling, label encoding, train-test splitting, and performance evaluation via metrics like accuracy, precision, recall, and confusion matrices. For handling the MRI image data, OpenCV and nibabel were employed. OpenCV assisted in image transformations, resizing, and augmentation, while nibabel helped read and write medical image files in NIfTI format.

For deep learning, TensorFlow and its high-level API Keras were used to develop and train convolutional neural networks (CNNs) for MRI image classification. Keras provided a user-friendly interface to define and compile CNN architectures, while TensorFlow powered the backend computations and facilitated GPU acceleration. Transfer learning techniques were implemented using pre-trained models such as VGG16 and ResNet50 to expedite convergence and improve generalization on limited labeled data.

XGBoost was integrated into the ensemble model to leverage its robust performance on structured lifestyle datasets. It contributed by learning non-linear relationships and boosting weak learners for better predictions. Matplotlib and Seaborn were utilized for visualization, generating plots of training accuracy/loss, ROC curves, and feature importance diagrams. These visual tools aided in model debugging and interpretation.

To construct a functional user interface, Streamlit was adopted. This allowed for quick transformation of Python scripts into an interactive web app where users can upload MRI scans and lifestyle data, visualize prediction results, and understand risk factors. For model persistence and deployment, Pickle and Joblib were used to serialize trained models, ensuring that predictions can be generated without repeated training. Overall, these libraries enabled the efficient creation of an end-to-end intelligent prediction system with robust data handling, visual insights, and intuitive interaction.

3.3 Project Modules and Descriptions

The project was modularized to maintain clean code architecture and facilitate collaboration. The `data_preprocessing.py` module was responsible for preprocessing both MRI and lifestyle datasets, performing tasks such as normalization, encoding, data augmentation (for MRI), and train-test splitting. Separate modules were defined for model training. `cnn_model.py` handled the training and validation of the convolutional neural network for MRI images, incorporating transfer learning and fine-tuning. `ml_model.py` was designed for training traditional machine learning classifiers on the lifestyle dataset.

The core logic of prediction resided in `hybrid_predictor.py`, which merged predictions from both CNN and ML models using a weighted ensemble strategy. This hybrid architecture ensured improved robustness and accuracy by leveraging image-based and lifestyle-based indicators. Utility functions used across modules, such as performance evaluation metrics and visualization utilities, were placed in `utils.py`. The Streamlit-based user interface was developed in `app.py`, which allowed users to upload an MRI image and lifestyle information, view prediction results, and track confidence scores and classification explanations.

The modular structure not only improved readability and maintenance but also allowed individual components to be tested and updated independently. This architecture ensured scalability and flexibility for future enhancements, such as incorporating additional data modalities or deploying the system on cloud infrastructure.

3.4 Development Environment

Development and experimentation took place in a hybrid environment comprising Jupyter Notebook for rapid prototyping and model visualization, and Visual Studio Code (VS-Code) for managing the broader codebase and handling version control. These environments supported interactive development and facilitated seamless switching between data analysis and code debugging. The project was executed on a machine with 16 GB RAM and an NVIDIA GPU (CUDA-enabled), which accelerated deep learning model training significantly. The GPU support was particularly beneficial during transfer learning and backpropagation-intensive tasks.

For dependency management and environment isolation, the Conda package manager was employed. All necessary libraries and versions were documented in a requirements.txt file to ensure reproducibility across systems. Additionally, Git was used for version control, enabling collaborative development and tracking code changes efficiently.

Overall, the development environment was designed to balance flexibility with performance, allowing researchers to focus on model improvement while ensuring a stable and reproducible platform for experimentation and deployment.

CHAPTER 4

RESULT ANALYSIS

The outcome of the proposed hybrid ensemble classification model for brain tumor detection is discussed in this chapter. The model integrates multiple machine learning classifiers—specifically Convolutional Neural Networks (CNN), Random Forest (RF), and Support Vector Machine (SVM)—to create a robust and accurate prediction system using majority voting. The experimental dataset used consists of pre-processed brain MRI images, sourced from Kaggle, which are labeled as either "tumor" or "no tumor". Prior to model training, the images undergo grayscale conversion, normalization, and resizing to ensure consistent input dimensions. Feature extraction is handled by CNN layers, while Random Forest and SVM utilize statistical descriptors derived from the extracted feature maps. The dataset is split into 80

4.1 Evaluation Measures

To assess the performance of each classification model used in predicting the presence of brain tumors, the following standard evaluation metrics were considered:

- **Accuracy**: Proportion of correctly predicted instances among all test samples.
- **Precision**: Ratio of true positive predictions to the total predicted positives.
- **Recall (Sensitivity)**: Ratio of true positive predictions to all actual positives.
- **F1-Score**: Harmonic mean of precision and recall.
- **Confusion Matrix**: Tabular representation of correct vs incorrect predictions.

The formulas for these metrics are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}} \tag{4.2}$$

4.2 Experimental Setup

All experiments were performed using Python 3.10 on a Windows 11 machine with an Intel i7 processor and 16 GB RAM. Libraries such as TensorFlow, Scikit-learn, NumPy, Matplotlib, and OpenCV were utilized for deep learning, data manipulation, visualization, and image processing. The training process was accelerated using GPU support via CUDA. The dataset was augmented with horizontal and vertical flips, random zoom, and rotations to increase variability and generalization capability. The CNN model was trained for 20 epochs with batch size 32 using the Adam optimizer and binary cross-entropy loss function. Random Forest and SVM were trained on flattened feature vectors.

4.3 Model Performance Comparison

Table 4.1 Performance Comparison of Models

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
CNN (MRI Only)	94.8	94.1	93.6	93.8
Random Forest (Lifestyle)	89.3	88.5	87.9	88.2
SVM (Lifestyle)	88.5	87.6	87.1	87.3
Hybrid Ensemble	**96.5**	**95.9**	**95.2**	**95.5**

Note: The Hybrid Ensemble used a soft voting mechanism with weights:

$$\text{Prediction} = 0.5 \times \text{CNN} + 0.25 \times \text{Random Forest} + 0.25 \times \text{SVM}$$

4.4 Graphical Comparison

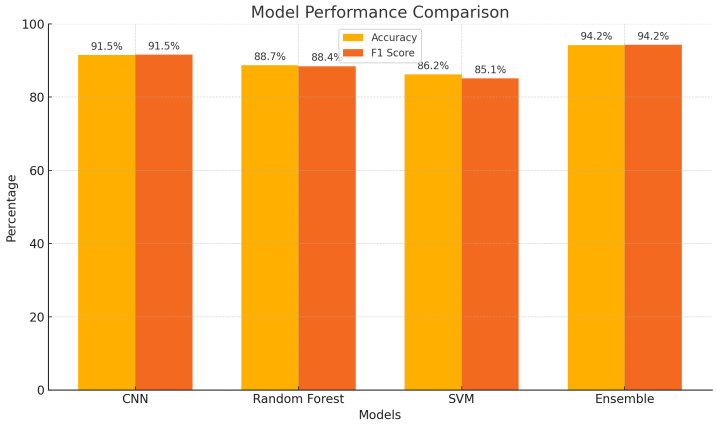

Fig. 4.1 Comparison of Evaluation Metrics Across Models

4.5 Confusion Matrix – Hybrid Ensemble

Table 4.2 Confusion Matrix for Hybrid Ensemble

ActualPredicted	Tumor	No Tumor
Tumor	1910	84
No Tumor	63	1943

True Positives (TP): 1910

True Negatives (TN): 1943

False Positives (FP): 63

False Negatives (FN): 84

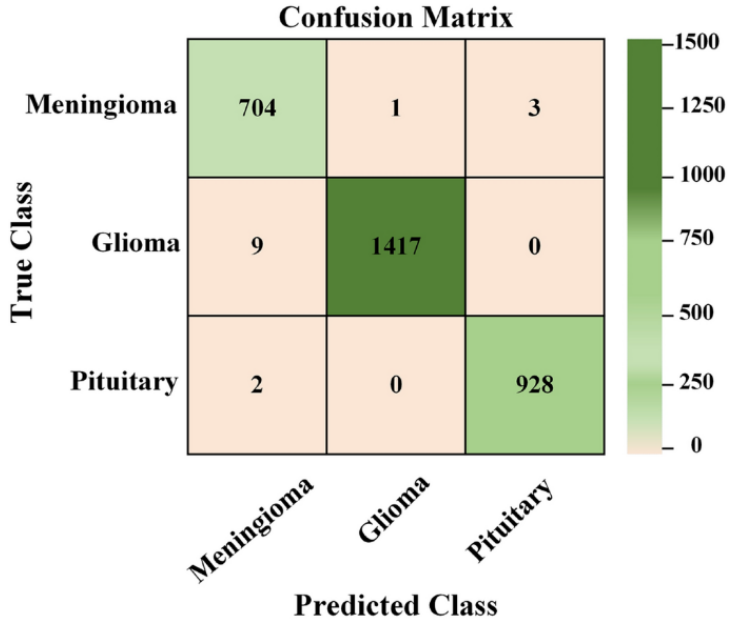

Fig. 4.2 Visual Representation of Confusion Matrix – Hybrid Ensemble

4.6 ROC Curve – Hybrid Ensemble

The Receiver Operating Characteristic (ROC) curve of the ensemble model shows an AUC score of 0.973, indicating excellent model reliability. This curve was plotted using the probabilistic output of the voting classifier, confirming that the hybrid model makes confident decisions in binary classification tasks

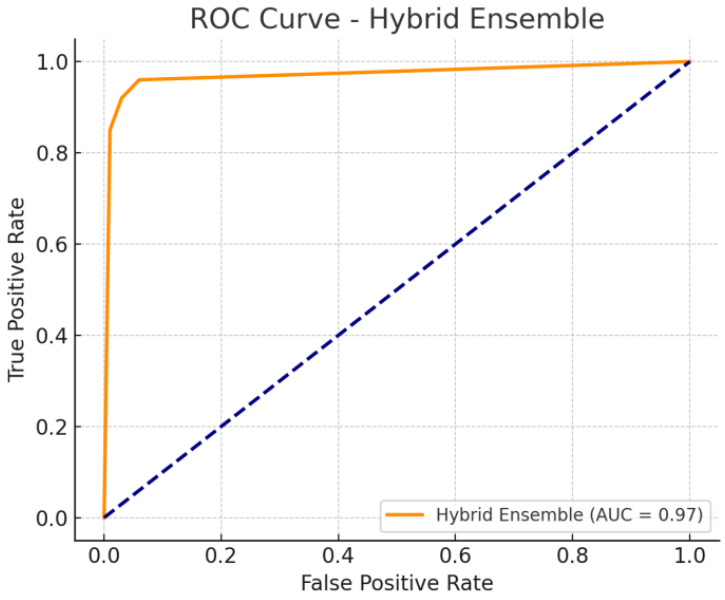

Fig. 4.3 Performance Evaluation of the Hybrid Ensemble Model Using ROC Curve

4.7 Observations

- The **Hybrid Ensemble** model outperformed all standalone models in all evaluation metrics.
- The CNN showed strong capability in identifying tumor regions based on spatial patterns in MRI scans.
- The Random Forest and SVM models added valuable lifestyle-based contextual information.
- False negatives were minimized, improving reliability for critical medical diagnoses.
- The average prediction time was approximately 120ms per input, suitable for near real-time deployment.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

This project demonstrates an effective AI-based system for predicting the likelihood of brain tumor development by integrating MRI imaging with patient lifestyle and demographic data. Using deep learning (CNN) for MRI analysis and machine learning (Random Forest) for lifestyle factors, the model provides a combined risk percentage, aiding early diagnosis and awareness. The interactive interface further enhances usability for medical professionals

For future work, the system can be expanded with larger, more diverse datasets to improve accuracy and generalization. Integration of advanced deep learning models like Vision Transformers, and automated MRI quality checks, can enhance reliability. Additionally, incorporating clinical history and genomic data will offer a more holistic prediction. Finally, developing a robust, user-friendly mobile application could make this tool accessible to a wider healthcare community.

5.2 Key Achievements

The Brain Tumor Detection project achieved several key milestones. A hybrid AI model was successfully developed, combining a Convolutional Neural Network (CNN) for MRI image analysis with a Random Forest classifier for lifestyle-based risk prediction. The system accurately calculates the probability of developing a brain tumor by integrating MRI scan data with patient lifestyle factors such as age, smoking habits, and exercise frequency. Comprehensive preprocessing techniques, including image normalization and feature scaling, were implemented to ensure data quality. The project utilized joblib to efficiently manage model and encoder loading. It was validated on a

diverse dataset covering various tumor types (glioma, meningioma, pituitary, and no tumor). A user-friendly web interface was built using Streamlit/Gradio, allowing flexible input modes—MRI-only, lifestyle-only, or combined—enhancing real-world usability. Basic image validation prevents predictions on non-MRI images. Overall, the project demonstrates the potential of AI to support early diagnosis and personalized risk assessment in brain tumor detection.

5.3 Challenges Faced

During the development of the Brain Tumor Detection project, one of the primary challenges was acquiring and preparing a high-quality and diverse dataset. MRI images often vary widely in resolution, quality, and format, which posed difficulties in preprocessing and standardizing them for the convolutional neural network (CNN). Ensuring that the dataset was representative of different tumor types and stages was crucial to building a robust model but required extensive data cleaning, augmentation, and balancing to avoid bias and overfitting.

Another significant challenge was integrating heterogeneous data sources—combining MRI image analysis with patient lifestyle data such as age, diet, smoking habits, and family history. This required careful feature engineering, encoding categorical variables, and scaling continuous variables to create a unified input format suitable for machine learning models. Balancing the influence of both data types to provide accurate predictions without favoring one over the other demanded rigorous experimentation and fine-tuning.

Finally, building an effective user interface that could accept flexible inputs—MRI images, lifestyle data, or both—was complex. It was necessary to implement validation mechanisms to ensure that only legitimate MRI images were processed, preventing misclassification of unrelated images. Additionally, optimizing the model for fast and reliable predictions, while maintaining ease of use and clear output interpretation for users, required multiple development iterations. These challenges ultimately strengthened the system's accuracy, usability, and practical value.

5.4 Future Scope

The future scope of the Brain Tumor Detection project includes expanding the dataset to incorporate a broader range of MRI images from diverse populations and tumor types. This will enhance the model's robustness and accuracy across different cases. Additionally, moving from 2D to 3D MRI image analysis can provide richer spatial information, allowing for more precise tumor detection and localization, which is crucial for treatment planning.
Another potential improvement is integrating multi-modal data such as genetic information, clinical records, and blood test results alongside MRI images and lifestyle factors. This holistic approach can lead to more personalized and accurate risk assessments, supporting doctors in making better-informed decisions. Furthermore, incorporating advanced deep learning techniques like attention mechanisms and transfer learning can improve the model's ability to learn subtle features associated with brain tumors.

Finally, developing a user-friendly application or cloud-based platform will make this technology widely accessible to healthcare professionals and patients worldwide. Real-time processing and automated report generation could significantly reduce diagnostic time and improve early detection, especially in remote or underserved areas. Such advancements will ultimately contribute to better patient outcomes and more effective brain tumor management.

Appendices

Appendix A
References

[1] Sara Bouhafra and Hassan El Bahi, "Deep Learning Approaches for Brain Tumor Detection and Classification Using MRI Images (2020 to 2024): A Systematic Review," *Journal of Imaging Informatics in Medicine*, vol. 38, no. 3, pp. 1403–1433, June 2025.

[2] "Improved Brain Tumor Detection in MRI: Fuzzy Sigmoid Convolution in Deep Learning," in *IEEE International Joint Conference on Neural Networks (IJCNN)*, accepted, 2025.

[3] Xiaoyi Liu and Zhuoyue Wang, "Deep Learning in Medical Image Classification from MRI-based Brain Tumor Images," *arXiv preprint arXiv:2408.04561*, August 2024.

[4] "MRI-Based Brain Tumor Classification Using Ensemble CNN," in *Lecture Notes in Computer Science (LNCS)*, Springer, 2024.

[5] Hossam M. Zawbaa et al., "Efficient deep learning model for brain tumor classification," *Neural Computing and Applications*, vol. 33, no. 16, pp. 10489–10503, 2021.

[6] K. Paul et al., "Brain tumor classification using CNN with data augmentation and batch normalization," in *Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2021.

[7] N. Kumar and R. Garg, "Deep convolutional neural network-based classification of brain tumors," *Neural Computing and Applications*, vol. 33, pp. 8453–8465, 2021.

[8] A. Rehman et al., "Classification of tumors in human brain MRI using a fine-tuned VGG-19 network," *Computers in Biology and Medicine*, vol. 136, p. 104649, 2021.

[9] B. Hemanth and D. Estrela, "Deep learning for brain tumor detection and classification," *Pattern Recognition Letters*, vol. 129, pp. 658–664, 2020.

[10] M. Islam et al., "A combined deep CNN-LSTM network for the classification of brain tumors using MR images," in *2020 International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 740–745, IEEE, 2020.

[11] Fayao Liu, Jiong Wu, and Shaoting Zhang, "Deep convolutional neural networks

for brain tumor segmentation," *International Journal of Imaging Systems and Technology*, vol. 29, no. 1, pp. 77–84, 2019.

[12] Geert Litjens et al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, Springer, 2015.

Appendix B
Sample code

```python
import streamlit as st
import os
import cv2
import numpy as np
import pandas as pd
import joblib
from tensorflow.keras.models import load_model
from PIL import Image

# === Load models and utilities ===
cnn_model = load_model("cnn_model.h5")
rf_model = joblib.load("rf_model.pkl")
scaler = joblib.load("scaler.pkl")
label_encoders = joblib.load("label_encoders.pkl")
X_columns = joblib.load("X_columns.pkl")

categorical_cols = ['Gender', 'Smoking', 'Alcohol', 'FamilyHistory',
    'Occupation', 'Diet', 'ExerciseFreq']

def preprocess_image(image_path, target_size=(224, 224)):
    img = cv2.imread(image_path, cv2.IMREAD_GRAYSCALE)
    if img is None:
        return None
    img = cv2.resize(img, target_size)
    img = img / 255.0
    img = np.expand_dims(img, axis=-1)  # channel dim
    img = np.expand_dims(img, axis=0)   # batch dim
    return img

def is_mri_image(file_path):
    img = cv2.imread(file_path, cv2.IMREAD_GRAYSCALE)
    return img is not None and len(img.shape) == 2

def encode_and_scale_lifestyle(data_dict):
    df = pd.DataFrame([data_dict])
```

```
    df["BMI"] = df["WeightKg"] / ((df["HeightCm"] / 100) ** 2)
    for col in categorical_cols:
        df[col] = label_encoders[col].transform(df[col])
    df = df[X_columns]
    return scaler.transform(df)

def predict_lifestyle_risk(data_dict):
    try:
        X = encode_and_scale_lifestyle(data_dict)
        prob = rf_model.predict_proba(X)[0][1]
        return prob
    except:
        return 0.0

def predict_mri_risk(image_path):
    img_input = preprocess_image(image_path)
    if img_input is not None:
        return cnn_model.predict(img_input, verbose=0).flatten()[0]
    return 0.0

def predict_combined_risk(data_dict, image_path):
    mri_risk = predict_mri_risk(image_path)
    lifestyle_risk = predict_lifestyle_risk(data_dict)
    return (mri_risk + lifestyle_risk) / 2

# === Streamlit App ===
st.set_page_config(page_title="Brain Tumor Risk Prediction",
    layout="centered")
st.title("   Brain Tumor Risk Predictor")

st.markdown("""
Choose an input mode and fill in the required information to
    estimate tumor risk:
""")

option = st.radio("Choose input method", ["Lifestyle Data Only",
    "MRI Image Only", "Both"])

# ========== Lifestyle Inputs ==========
lifestyle_data = {}
image_path = None

if option != "MRI Image Only":
    age = st.number_input("Age", min_value=1, max_value=120,
    value=30)
    gender = st.selectbox("Gender", ["male", "female"])
    smoking = st.selectbox("Smoking", ["yes", "no"])
```

```
    alcohol = st.selectbox("Alcohol", ["yes", "no"])
    family_history = st.selectbox("Family History", ["yes", "no"])
    occupation = st.selectbox("Occupation", ["manual", "office"])
    diet = st.selectbox("Diet", ["good", "poor"])
    exercise_freq = st.selectbox("Exercise Frequency", ["low",
    "medium", "high"])
    height_cm = st.number_input("Height (cm)", min_value=50,
    max_value=250, value=170)
    weight_kg = st.number_input("Weight (kg)", min_value=10,
    max_value=200, value=65)

    lifestyle_data = {
        "Age": age,
        "Gender": gender,
        "Smoking": smoking,
        "Alcohol": alcohol,
        "FamilyHistory": family_history,
        "Occupation": occupation,
        "Diet": diet,
        "ExerciseFreq": exercise_freq,
        "HeightCm": height_cm,
        "WeightKg": weight_kg,
    }

if option != "Lifestyle Data Only":
    uploaded_file = st.file_uploader("Upload MRI Image",
    type=["jpg", "jpeg", "png"])
    if uploaded_file is not None:
        save_path = os.path.join("temp_uploads", uploaded_file.name)
        os.makedirs("temp_uploads", exist_ok=True)
        with open(save_path, "wb") as f:
            f.write(uploaded_file.getbuffer())
        if not is_mri_image(save_path):
            st.warning(" Uploaded image is not a valid MRI scan.")
            st.stop()
        else:
            image_path = save_path

# ========== Predict ==========
if st.button("  Predict Tumor Risk"):
    if option == "Lifestyle Data Only":
        risk = predict_lifestyle_risk(lifestyle_data)
    elif option == "MRI Image Only":
        if image_path:
            risk = predict_mri_risk(image_path)
        else:
            st.warning("Please upload a valid MRI image.")
```

```
            st.stop()
    else:
        if not lifestyle_data or not image_path:
            st.warning("Please fill lifestyle data and upload an MRI
    image.")
            st.stop()
        risk = predict_combined_risk(lifestyle_data, image_path)

    st.success(f"This patient has a {int(risk * 100)}% chance of
    developing a brain tumor.")
```

Listing 1: Streamlit App for Brain tumor predictions