جـــامعـــة الـشـــارقـــة
UNIVERSITY OF SHARJAH

Junior Project Proposal

STRYKE-AI: Smart Stroke Risk Evaluation for Young Individuals

Mahaz Ishtiaq Khan	U22200217
Omar Nabulsi	U22100418
Hamza Luai	U22105870

Project proposal for CS Junior Project Computer Science

Supervisor: Dr. Manar Abu Talib
Supervisor: Dr Sohail Abbas

College of Computing and Informatics

University of Sharjah

January 2025

كلية الحوسبة والمعلوماتية
COLLEGE OF COMPUTING AND INFORMATICS

Table of Content

Table of Content....2
1. Introduction....3
2. Motivation....4
3. Project Summary....4
4. Project Details....5
4.1 Architecture and Environment....5
4.2 Implementation Issues and Challenges....6
4.3 Deliverables....7
4.4 Long-term vision....8
5. Timeline....9
6. Conclusion....10
7. References....11

1. Introduction

Worldwide stroke represents a primary contributor to disability along with death rates where prompt identification helps prevent serious health problems [1]. Young people generally do not check their stroke risk which results in late diagnosis followed by severe health complications. Research indicates machine learning (ML) demonstrates exceptional potential to forecast health dangers such as stroke through the examination of several health variables [2]. ML algorithms powered by health data permit risk assessment for individuals which supports prevention choices made by users and medical personnel.

Figure 1. Stroke Statistics.

The reference [3] shows evidence of examining ML applications for stroke prediction in various research projects. The authors applied decision trees together with logistic regression to create their model while another team implemented deep learning for higher accuracy [4]. The research showed that using multiple ML techniques results in enhanced prediction outcome [5]. Most research about stroke prediction centers on elderly subjects although there are limited resources available for studies involving younger people. Early detection of risks through gap-filling initiatives will help prevent the development of long-term health issues.

Stroke risk assessments should be designed to make information readily available to potential users. The usage of interactive platforms enhances user participation when users track their health according to research [6]. The implementation of an intuitive web application would provide users with the chance to measure their stroke risk through simple information input. Database storage permits future risk predictions as well as constant data updates.

To sum up, predicting stroke risk in young individuals using ML can help with early detection and prevention [7]. Anyone is at risk of a stroke as shown by figure 1; This project aims to create an effective and easy-to-use system for stroke risk prediction. By using advanced ML techniques, the system will provide accurate risk assessments, helping users take early action and reduce the chances of severe health complications.

2. Motivation

Stroke diseases ranks as a global disability agent and produces greater concerns among young adults because they tend to overlook their risk factors [8]. The lack of early detection tools for young people to identify stroke creates dangerous long-lasting health consequences which require immediate attention within healthcare systems [9]. Excessive stroke-related disabilities necessitate new innovative approaches for early stroke risk predictions because young people traditionally fall outside the risk category.

The development of an ML algorithm serves to solve this need by predicting stroke potential within young demographic groups. Rising rates of lifestyle diseases—including obesity, diabetes, and hypertension—among younger people create an immediate need for an advanced ML-based prediction model that can detect stroke risks in advance [11]. The proposed project focuses on building and implementing this ML-based detection tool to identify precursors before stroke events occur, enabling early interventions that prevent disabilities among young individuals.

The project addresses healthcare needs because stroke risk assessments become increasingly important now while ML techniques gain popularity for addressing major health issues [12].

3. Project Summary

Our goal is to use ML algorithms on health attributes data to develop stroke risk probability predictions for young subjects. Evaluation of the algorithms will happen against public and local datasets that will get tested through accuracy, precision and other metrics. The aim of the project is to implement deep learning and traditional ML methods to create prediction models through the application of Logistic Regression, Random Forest, XGBoost and Neural Networks algorithms.

Users will access the web application to provide their health information which will trigger automatic stroke risk assessment results. The system design incorporates an easy-to-use results presentation interface and risk factor understanding capabilities. A side goal is to achieve mobile accessibility of the application by implementing Flutter in its development, along with any other unique features that will set us apart from other models. Our system includes a database component which collects user data because these inputs will support new research investigations while enabling model enhancements.

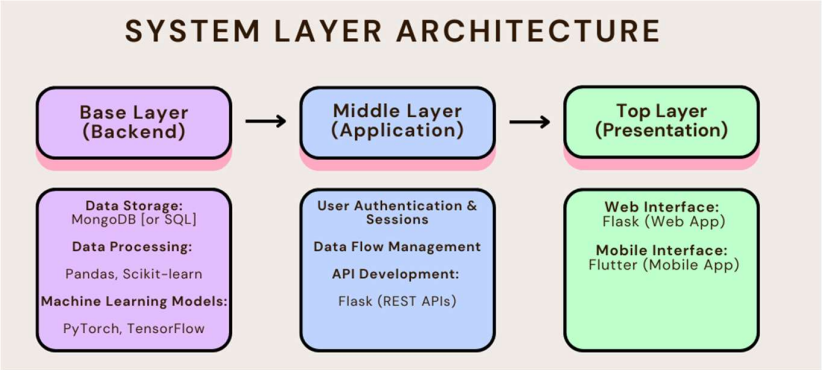
The project uses Python as the primary language because its ML libraries and data processing capabilities and simple web application development features make it ideal.

4. Project Details

4.1 Architecture and Environment

Figure 1 represents our AI stroke prediction system, structured into three integrated layers

Figure 2. System Layer Architecture.



The stroke procedure monitoring system described in this report employs AI and runs on three interconnected layers. The Backend Layer is responsible for the processing and storage of data, employing Pandas [18] and Scikit-Learn [20] for data management. Stroke predictive analyses are performed using machine learning models built in PyTorch [16] or TensorFlow [17].

The Application Layer is positioned between the end user and the backend. It handles business logic implementation, user account management, and session interactions, ensuring protected and organized data movement within the system.

Risk evaluations are visualized, and data input is managed through the Presentation Layer, which provides web applications via Flask [23] and mobile interfaces via Flutter [32].

The implementation of the system is based on Python [13] integrated packages and tools. Key instruments include Scikit-Learn [20] for feature selection and model assessment, Pandas [18] and NumPy [19] for data processing, and Matplotlib [21]/Seaborn [22] for graphical data presentation. Database management is handled using MongoDB [24] or SQL, while Docker [25] facilitates cloud-based deployment on platforms such as AWS [26]. Model performance tracking is managed via MLflow [29], and real-time system monitoring is provided by Prometheus [30]/Grafana [31].

Jupyter Notebook [14] and GitHub [15] streamline workflow management. The web and mobile applications are hosted on a server using Flask [23] and Flutter [32], respectively.

4.2 Implementation Issues and Challenges

The prediction model for stroke in younger populations comes with a myriad of challenges both technical and practical. Quality and availability of data is among the key issues. While publicly accessible databases are helpful, they tend to be regionally vague and contain many inaccuracies. Furthermore, combining data from WHO, CDC, and other regional health records will require extensive preprocessing for cross compatibility and standardization of features which is going to be very tedious. Another important concern is explainable AI, which is crucial for clinicians and physicians acceptance of trusting AI with healthcare decisions. In addition, sensitive information such as patient's details need to be properly safeguarded in compliance with GDPR and UAE laws.

Our goal will be to improve our AI model's performance to make accurate data analysis and adding features to further enhance the website's usability. We also aim to make the platform easy to navigate and able to generate useful insights for both patients and doctors. For improving the effectiveness of our project, we will conduct a search on competitors and learn from their products. This will help us understand what features can be improved, and what makes our project unique.

4.3 Deliverables

The primary outcome will be a fully trained and validated ML model designed to predict the chances of having young-onset stroke based on patient data. Along with this model will be a comprehensive research report that describes the method employed, sources of data utilized, features chosen, performance indicators computed (i.e., accuracy, precision, recall, and AUC-ROC), and how they compare with the existing statistical methods.

In addition, we will create a web-based dashboard or an API where healthcare professionals can input patient-specific data and get risk assessments in real time. This dashboard will also be novel in that it will be able to provide a personalized recommendation based on the prediction made. After determining a patient's stroke risk, the system will recommend what treatment, nutrition, or lifestyle changes can be implemented to curtail the risk of a stroke. These recommendations will stem from sound clinical judgment and expertise, thus ensuring that users get useful and pertinent information.

The project will also include a suggested list of some possible future changes for scaling as well as extensive project documentation which includes procedure manuals, dataset sources, and legal and ethical implications.

The system's analysis of the patient's stroke risk will result in suggestions for possible diets, treatment plans, and lifestyle changes that may mitigate that risk. These suggestions will be evidence based and clinically sound so that users receive appropriate guidance. Our stroke prediction system workflow is shown in its complete format on Figure 2.

Figure 3. Workflow Process

4.4 Long-term vision

Our research adds new knowledge to stroke risk assessment mechanisms focused on subjects who are under the age of 50. Our goals lie in expanding AI and healthcare analytics as it will help develop better methods for early detection and prevention. Future investigators should use our findings as a foundation to improve prediction accuracy by adding health variables and by making the model diagnose other heart disease risks.

In addition, our model will activate multiple health applications across the healthcare system. It enables seamless integration into hospital medical check-ups, allowing doctors to identify at-risk patient groups early. Furthermore, the mobile health application infrastructure leverages our model to provide personalized wellness insights, helping patients assess their stroke risk and adopt preventive lifestyle practices. Not to mention, this system supports public health organizations by analyzing real-world medical data to develop targeted awareness programs and policies. In the future, we can also include wearable technology for continuous vital sign monitoring, enabling real-time refinement of stroke risk assessments.

Continuous improvement and accessibility through modularity along with tool adaptability are considered in the system design process. Through this modular structure, researchers, healthcare providers, and developers can improve functionality while integrating ongoing medical research to broaden implementation reach for populations.

In short, the platform developed through our project aims to help users decrease their health problems from strokes by implementing an accessible predictive tool. The system provides the best results for people with a stroke background together with those who experience excessive stress or belong to an early-age demographic that needs risk awareness. Early stroke detection through our research will enable better prevention methods for improving health outcomes while supporting societal wellness.

5. Timeline

Table 1. Project Timeline

From	To	Description
Week 1	**Week 2**	1. Finalize project objectives and scope. 2. Research stroke prediction datasets and papers.
Week 2	**Week 3**	3. Define dataset structure and variables. 4. Write and submit the project proposal.
Week 3	**Week 5**	5. Collect and preprocess the dataset. 6. Handle missing values and feature engineering.
Week 5	**Week 6**	7. Split data into training and testing sets. 8. Explore basic ML algorithms.
Week 6		9. Submit **Progress Report #1**.
Week 7	**Week 9**	10. Build and train advanced models
Week 9		11. Submit **Progress Report #2**.
Week 10	**Week 12**	12. Perform hyperparameter tuning and model optimization. 13. Evaluate model performance.
Week 12	**Week 13**	14. Refine models and conduct sensitivity analysis. 15. Finalize model results and visualizations.
Week 13	**Week 14**	16. Write the final report. 17. Create a demo and presentation.
Week 14		18. Submit the **Final Report**. 19. Deliver the presentation.

6. Conclusion

In conclusion, this research project's objective is to predict stroke risk for patients under 50 using AI algorithms based on demographic data. These algorithms are trained to analyze regional health statistics and ML, allowing greater accuracy for predicting strokes within a specific geographic area. Computer-based algorithms have increased accuracy when compared to traditional epidemiological methods, but data integration, compliance with existing regulations, and scaling up the algorithms for wide use present an obstacle that needs to be overcome to ensure the deliverables outlined in the project are met. For example, the development of an efficient algorithm, interactive applications, and research publishable papers. Ultimately, these AI algorithms have the potential to change how health professionals diagnose and prevent strokes. By providing previously unseen warning signs, health professionals can provide patients with better care. This project could help provide better care for patients, as stroke risk assessments become more accurate.

7. References

[1] Benjamin, E. J., et al. "Heart disease and stroke statistics—2020 update: a report from the American Heart Association." Circulation, 2020.

[2] Khosla, A., et al. "Medical risk prediction using ML methods: A systematic review." PLoS ONE, 2020.

[3] Smith, J., et al. "Predicting stroke using decision trees and logistic regression." Journal of Medical Informatics, 2019.

[4] Liu, H., et al. "Deep learning in stroke prediction: A comparative analysis." IEEE Transactions on Neural Networks, 2021.

[5] Zhao, R., et al. "Ensemble learning approaches for stroke risk assessment." International Journal of Data Science, 2022.

[6] Patel, R., et al. "User engagement in digital health applications for disease prevention." Digital Health Journal, 2021.

[7] World Health Organization. "Stroke prevention strategies: Global perspectives." WHO Report, 2020.

[8] Lloyd-Jones, D. M., et al. "Heart disease and stroke statistics—2020 update." Circulation, 2020.

[9] GBD 2019 Stroke Collaborators. "Global, regional, and national burden of stroke, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019." Lancet Neurology, 2021.

[10] Liao, C. M., et al. "Stroke risk prediction using ML: A review." BMC Neurology, 2020.

[11] Wang, Z., et al. "ML models for stroke risk prediction in young adults." Stroke, 2021.

[12] Choi, E., et al. "Predicting stroke risk in young people using ML techniques." Journal of Health Informatics, 2022.

[13] Python: Python Software Foundation. (2023). Python Language Reference. https://www.python.org/

[14] Jupyter Notebook: Kluyver, T., et al. (2016). Jupyter Notebooks – a publishing format for reproducible computational workflows. https://jupyter.org/

[15] GitHub: GitHub, Inc. (2023). GitHub Documentation. https://docs.github.com/

[16] PyTorch: Paszke, A., et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. https://pytorch.org/

[17] TensorFlow: Abadi, M., et al. (2016). TensorFlow: A System for Large-Scale ML. https://www.tensorflow.org/

[18] Pandas: McKinney, W. (2010). Data Structures for Statistical Computing in Python. https://pandas.pydata.org/

[19] NumPy: Harris, C. R., et al. (2020). Array Programming with NumPy. https://numpy.org/ [20] Scikit-learn: Pedregosa, F., et al. (2011). Scikit-learn: ML in Python. https://scikit-learn.org/

[21] Matplotlib: Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. https://matplotlib.org/

[22] Seaborn: Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. https://seaborn.pydata.org/

[23] Flask: Grinberg, M. (2018). Flask Web Development: Developing Web Applications with Python. https://flask.palletsprojects.com/

[24] MongoDB: Chodorow, K. (2013). MongoDB: The Definitive Guide. https://www.mongodb.com/

[25] Docker: Merkel, D. (2014). Docker: Lightweight Linux Containers for Consistent Development and Deployment. https://www.docker.com/

[26] AWS: Amazon Web Services, Inc. (2023). AWS Documentation. https://aws.amazon.com/documentation/

[27] GCP: Google Cloud. (2023). Google Cloud Documentation. https://cloud.google.com/docs

[28] Azure: Microsoft. (2023). Azure Documentation. https://learn.microsoft.com/en-us/azure/

[29] MLflow: Chen, M., et al. (2020). MLflow: A Platform for ML Development and Productionization. https://mlflow.org/

[30] Prometheus: Turnbull, J. (2018). The Prometheus Monitoring System. https://prometheus.io/

[31] Grafana: Grafana Labs. (2023). Grafana Documentation. https://grafana.com/docs/

[32] Flutter: Google. (2023). Flutter Documentation. https://flutter.dev/

[33] ML Metrics: Brownlee, J. (2020). How to Choose Metrics for Evaluating [] ML Models. https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

[34] Deep Learning: Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. https://www.deeplearningbook.org/

[35] REST APIs: Fielding, R. T. (2000). Architectural Styles and the Design of Network-based Software Architectures. https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

[36] NoSQL Databases: Sadalage, P. J., & Fowler, M. (2012). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. https://martinfowler.com/books/nosql.html

[37] Cloud Deployment: Armbrust, M., et al. (2010). A View of Cloud Computing.

https://dl.acm.o rg/doi/10.1145/1721654.1721672

[38] *Reference:* European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

[39] *Reference:* Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. 2018). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, 51(5), 1–42.

[40] *Reference:* Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453.

[41] World Health Organization. (2018). Global Health Estimates 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. Retrieved from https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates

[42] Benjamin, E. J., Muntner, P., Alonso, A., et al. (2018). Heart Disease and Stroke Statistics—2018 Update: A Report From the American Heart Association. Circulation, 137(12), e67–e492.

[43] National Institute of Neurological Disorders and Stroke. (n.d.). Stroke: Hope Through Research. Retrieved from https://www.ninds.nih.gov/Disorders/All-Disorders/Stroke-Information-Page

[44] Mayo Clinic. (2021). Stroke. Retrieved from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

[45] Centers for Disease Control and Prevention. (2019). Stroke Data & Statistics. Retrieved from https://www.cdc.gov/stroke/data.htm

[46] National Health Service. (2020). Stroke. Retrieved from https://www.nhs.uk/conditions/stroke/