National College of Ireland

Optimizing Healthcare Framework using Cognitive Computing Techniques in Cloud: A Study on Enhancing Diagnostic Accuracy and Decision-Making

MSc Research Project
Cloud Computing

Rajaram Jagadeeswaran

Student ID: x22239243

School of Computing
National College of Ireland

Supervisor: Jitendra Kumar Sharma

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Rajaram Jagadeeswaran
Student ID:	X22239243
Programme:	Masters in Cloud Computing
Year:	2024
Module:	MSc Research Project
Supervisor:	Jitendra Kumar Sharma
Submission Due Date:	16/09/2024
Project Title:	Optimizing Healthcare Framework using Cognitive Computing Techniques in Cloud: A Study on Enhancing Diagnostic Accuracy and Decision-Making
Word Count:	8086
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Rajaram Jagadeeswaran
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	Q
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	Q
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	Q

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Healthcare Framework using Cognitive Computing Techniques in Cloud: A Study on Enhancing Diagnostic Accuracy and Decision-Making

Rajaram Jagadeeswaran
x22239243
MSCCLOUDB – MSc Research Project
National College of Ireland, Dublin

Abstract

The utilization of cloud platforms with cognitive technologies in healthcare delivers significant enhancements in decision-making and diseases diagnostics. The objective of this research project aims to address present difficulties in the using cognitive technologies in healthcare by utilizing cloud platforms features such as scalability, processing capability, and productive storage. The primary objective of this research is to investigate the possible breakthrough potential of cognitive computing in healthcare while simultaneously being aware of its limitations. Through a comparative evaluation, this study will examine the efficiency of various cognitive platforms such as generative AI's foundation models from AWS Bedrock and AWS SageMaker in disease prediction tasks. We proposed utilizing real-world datasets associated with specific diseases, with a focus on Text Generation to evaluate the diagnostic decision-making efficiency and accuracy of these platforms. The procedure incorporates collecting data, training the model, fine-tuning, deployment, and extensively evaluating its accuracy. Using AWS SageMaker for custom model deployment and AWS Bedrock for leveraging pre-trained models, we will fine-tune and deploy these models, followed by a comprehensive benchmarking process. Measures of performance including accuracy, precision, and inference times will be examined by the comparison framework. The evaluations from the investigations and implementations are intended to advance the development of cognitive computing technologies with the benefits and drawbacks in healthcare and provide insightful information for further study and the development of prototype in this domain.

Keywords: Healthcare, Cognitive Computing, Cloud Computing, Diagnostic Accuracy, Decision-Making, Comparison Framework, AWS Bedrock, AWS SageMaker, Streamlit.

1. Introduction

A notable advancement in healthcare has resulted from the integration of cloud infrastructure and cognitive computing over the years. AI and machine learning-based cognitive computing systems have the potential to revolutionize healthcare diagnosis as well as decision-making processes by providing more accurate, efficient, and data-driven insights. Better outcomes can be achieved by combining these cognitive systems with the scalability and computational capability of cloud platforms, which can handle massive

volumes of health care data and result in more effective storage with reduced processing times. The industry of healthcare technology is evolving rapidly, and there is a growing emphasis on using these developments to optimize clinical operations and enhance patient outcomes.

1.1 Background

Cognitive computing can be understood as a model wherever advanced information technology & artificial intelligence (AI) are combined to simulate the thought processes of humans. Finally, this type of technology analyses massive amounts of data in search of patterns and insights using various AI techniques, such as deep learning, machine learning, natural language processing, and so on. For this reason, cognitive computing has the potential to fundamentally alter how professionals diagnose diseases, develop drugs, and how patients are treated in the field of healthcare. Cognitive systems might help in the analysis of complex medical data to provide more precise diagnoses, patient outcomes predictions, and customized treatment protocols. The market for cognitive computing has been expanding quickly over the past few years, as seen by Figure 1 below. The upsurge is a result of increased research and development efforts in cognitive computing technologies, which are essential to many industries, including the provision of healthcare services. This is a result of the growing usage of cognitive computing-based advanced healthcare optimization solutions, which shows a trend toward increasingly sophisticated applications.

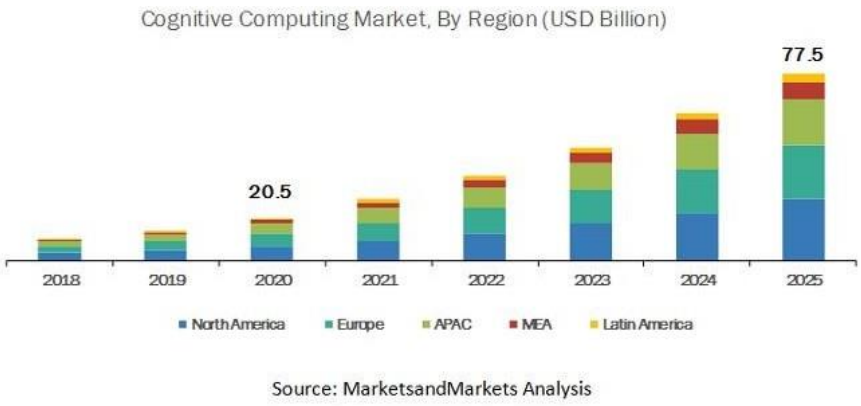

Figure 1: Market Trends in Cognitive Computing

But conventional data analysis techniques frequently find it difficult to keep up with this overflow of data, which may result in mistakes or inefficiencies in patient care. To overcome these challenges, it has developed sophisticated algorithms that are faster than human capabilities at analysing and understanding enormous amounts of data. These systems can provide individuals with more effective treatment and better results because they can discover hidden insights, identify early disease indications, and provide advice based on the data acquired from these types of results. Even so, cognitive computing is generally underutilized in the healthcare industry, despite its enormous potential to revolutionize the current delivery of healthcare services. It is essential that the huge influx of medical data be managed economically. Furthermore, to create models, express predictions, evaluate them and validating models, these systems need to regularly learn from health data. To ensure that the systems produce accurate outcomes and that all users may depend on them,

ethical concerns about data security and privacy must also be taken into consideration. So, until now, there appears to be an important hurdle to the integration of artificial intelligence in the healthcare industry.

1.2 Problem Statement

As one of the most common causes of mortality and permanent disability worldwide, stroke necessitates immediate medical attention to mitigate its effects and enhance the results for patients. However, determining the risk(s) of stroke is a complicated process that depends on a variety of variables, including past medical history, current lifestyle, and habits, any family history of stroke, and unpredictable information on health. That's why I have chosen this specific disease to explore in Text Generation experiments on various platforms to measure accuracy. While AWS SageMaker is a comprehensive machine learning service that provides developers with the ability to build, train, and deploy ML models at scale and AWS Bedrock, its focus on generative AI, offers a different approach by leveraging foundation models to generate text-based outputs. Understanding stroke prediction through these two platforms helps us identify their strengths and weaknesses as well as how they can be used together to improve healthcare outcomes. Despite having high intentions for cognitive computing in healthcare, there is still plenty of opportunity for improvement in how it's used. Among other issues, the scalability, accuracy, and compatibility of the current methods with the health systems are significant problem. Furthermore, not enough research has been done on the effectiveness of various cognitive computing platforms in particular medical tasks like disease prediction. This should encourage further research focused on handing in the gaps so that we can fully capitalize on the revolutionary potential of cognitive computing within our healthcare framework.

1.3 Research Objectives

This research aims to explore the transformative potential of cognitive computing in healthcare by evaluating the effectiveness of specific comparison on AWS SageMaker & Bedrock in disease prediction tasks. The primary objectives are:

- To develop and deploy a stroke prediction model using AWS SageMaker with real-time datasets. Utilizing AWS Bedrock to create a text generation model that can generate results based on stroke diagnosis prediction.
- To conduct a comparative analysis of these platforms AWS SageMaker and AWS Bedrock by focusing on stroke prediction (text generation).
- To identify the strengths and weaknesses of cognitive computing applications in healthcare, providing insights for future improvements.

Research Question. "To what extent does the implementation of cognitive computing techniques in the cloud computing platform impact disease diagnostic processes in terms of accuracy and efficiency?"

By illustrating how cutting-edge cognitive computing techniques may improve the predictive accuracy and interpretability of stroke prediction models by understanding these objectives, this project aims to advance disease diagnosis and its accuracy. We have included cloud computing into our work to guarantee the safe and effective management of massive data collections. We can develop an extensive framework for the application of AI in clinical settings primarily to the combined strength of AWS SageMaker and Bedrock.

Our comparison's results will show the potential of these technologies as well as offer suggestions for future advancements around AI-driven healthcare solutions.

2. Literature Review

The following section discusses current study which utilize cloud-based cognitive computing to enhance health care systems. To determine findings and areas in need of more research, we will thoroughly evaluate and contrast the various works that have been connected to this investigation. The understanding of the more general study project will improve when we know how these works relate to one another. Advances in artificial intelligence (AI) technologies in the medical industry have significantly impacted several businesses. Major multinational corporation (MNC) stakeholders are given access to cloud platforms that can be leveraged to improve real-world healthcare services, as well as insights into types of cognitive computing for their enterprises.

Intelligent diagnosis techniques and tailored treatment plans have been made possible by the integration of AI and cloud machine learning expertise. According to one study, in addition to attaining precision medicine, CDSS is one method for enhancing diagnosis accuracy (Castaneda et al., 2021). These platforms process massive amounts of clinical data using clouds, providing instant insight and guidance to healthcare providers. To acquire the trust of both patients and healthcare practitioners, data integrity and confidentiality must be guaranteed, as stated by Castaneda et al. (2021). AWS SageMaker was used by Beragu, Suraj. (2022) to achieve architecture on the E-Healthcare system. In this related work, CNN & K-NN Classifier with Random Forest algorithms were used to predict diseases; the accuracy rates of each algorithm were compared with the rates of the others. Furthermore, this study included the fog computing paradigm, which is an additional notable feature.

2.1 Cognitive Computing in Healthcare:

In the healthcare industry, cognitive computing is a new style of thinking that processes and analyzes enormous quantities of medical data using artificial intelligence (AI). The application of cognitive computing in healthcare has the potential to significantly enhance patient outcomes, personalize treatment plans, and increase the accuracy of diagnosis. Numerous research has been done on this subject, highlighting both the advantages and difficulties encountered during the implementation process. This technology has several benefits, one of which is its capacity to handle unstructured data. According to Davenport and Kalakota (2019), cognitive systems could read materials such as patient records, studies of medical literature, and results from clinical trials to provide information that can be utilized to clinical decision-making. When handling complex diagnoses where conventional procedures are ineffective, this characteristic becomes crucial. On the other hand, there are certain drawbacks to integrating cognitive computing in a healthcare environment, primarily related to privacy issues with patient data security precautions that shouldn't be disregarded. supervision of authenticity, confidentiality, and availability. According to Bhardwaj & Sharma (2020) non-disclosure can never be compromised when it comes to the architecture of the systems that maintain people's health records. Ethical problems must be resolved before decisions on the usage of such technologies are made to win over the trust of medical professionals and their patient base.

Healthcare organizations need to invest in supercomputers and train their staff on how to operate them effectively, as noted by Jiang et al. (2017). Behera, Bala, and Dhir (2019) carry out a comprehensive analysis of the literature pertaining to healthcare and cognitive computing. They talk about how cognitive systems might improve decision-making and diagnostic accuracy, but they also point out several drawbacks, such as poor data quality, privacy difficulties, and a lack of defined procedures for data integration and analysis. Like this, Sharma and Ghosh's (2022) research focus on the potential for improving patient outcomes using cognitive platforms in healthcare systems. When employing such a system in clinical practice, they bring up ethical concerns about privacy. As a result, they call for these issues to be solved if the medical community is to accept the system more widely. Alowais et al. (2023) provide a thorough explanation of the application of artificial intelligence in clinical practice, demonstrating how AI-enabled decision support systems enhance patient outcomes and enhance the quality of care provided by healthcare institutions. The report also discusses the difficulties with data privacy in this new technological era, when it is simple to expose personal data if appropriate security measures aren't put in place. Training medical professionals on the many AI tools available for use during their daily tasks linked to patient care management is another topic that the authors highlight. Srivani et al. (2023) Examines current and upcoming developments in the field of cognitive computing for healthcare, including the use of deep learning techniques in many industries and the integration of electronic health records (EHRs) among other things. The significance of protecting people's personal information and ethical implications of these applications is also addressed. Bhuiyan and Islam, (2023) provide a scalable cloud-based healthcare system management framework. They contend that the only way to accomplish scalability, availability, and sustainability is to combine cloud technologies with cognitive computing to create an architecture that can manage massive volumes of data sets at any given moment. It is best to leverage cloud-based technologies for real-time data processing and decision making to diagnose patients accurately and efficiently. In Davenport and Kalakota (2019) study, explore the possible usages advantages, and disadvantages of artificial intelligence in the healthcare industry. The authors give an outline of how AI can change the way healthcare is delivered and enhance patient outcomes, emphasizing patient management and diagnostic procedures.

2.2 Cloud Computing in Healthcare:

Cloud computing has completely changed the healthcare industry by providing a flexible, affordable, and effective data processing, analysis, and storage system. The use of cloud computing in healthcare allows for the integration of many information sources, resulting in extensive data analysis and improved patient care. Cloud-based frameworks facilitate real-time data processing, which is essential for timely and accurate clinical decision making (Islam & Bhuiyan, 2023). Building physical infrastructure that can handle massive amounts of data and be scaled up or down according to requirements without the assistance of cloud services would be too expensive for health care providers. Cloud systems include extra security elements in addition to guaranteeing privacy protection for private medical documents. Beragu (2022) claims that the efficiency and accuracy of disease prediction tools greatly increase when they are integrated into patient monitoring systems since this software enhances their functionality. Cloud computing enhances the capabilities of healthcare systems by offering real-time analytics and continuous patient health tracking when combined with Internet of Things sensors for analysis.

2.3 Evaluation of Text Generation - Stroke Prediction Models in Disease Diagnostics*:*

With the statistics, it’s seen as a crucial area of healthcare given the morbidity and fatality rates associated with stroke. A variety of models and methods have been developed for predicting the risk of having a stroke and they include newly developed models, sophisticated machine learning algorithms as well as more conventional statistics-based methods. For example, classic models such as the Framingham Stroke Risk Profile employ logistic regression to assess the possibility of being affected based on clinical and demographic characteristics (Wolf et al., 1991). However, they frequently fall short because of the way they are designed, which prevents them from accounting for intricate relationships between various variables. Machine learning is another technique that has shown to be successful in raising prediction accuracy; it makes use of computer systems to automatically identify patterns in massive datasets. According to Yang et al. (2018) research, machine learning techniques such as support vector machines, random forests, and neural networks can identify patterns that would otherwise go undetected if traditional models were the only ones used. Furthermore, these tools can integrate several kinds of data, including genetic, imaging, and clinical records, among others, to provide a comprehensive risk assessment. Deep learning models, in particular convolutional neural networks (CNNs), can also be used to evaluate medical images to identify stroke symptoms early on. An example of this is the study conducted by Kamnitsas et al. (2017), who stated that by segmenting brain lesions effectively using an MRI scan in conjunction with their proposed DL model, early intervention was possible.

Huang, X., et al. (2022) provided novel perspectives on applying machine learning to develop stroke prediction models in hypertensive individuals. It was discovered that a variety of machine learning algorithms could accurately forecast the risk of stroke, and that's why it's critical to have models tailored to certain patient groups. Kamnitsas et al. (2017) proposed an effective multi-scale 3D convolutional neural network (CNN) with fully linked conditional random fields (CRF) for accurate brain lesion segmentation. While the focus of the presentation was brain lesions, similar techniques can also be applied to the development of more sophisticated stroke prediction models when working with imaging data. Jiang et al. (2017) explore artificial intelligence in healthcare from past to future history with especially focused on stroke and vascular neurology. The researchers also provide an overview of AI-based techniques for stroke diagnosis and prediction, tracking the evolution of these technologies over time and outlining possible directions for future improvement. Using AWS SageMaker, Beragu (2022) developed an e-healthcare system in which several algorithms were used to forecast diseases. This provides insight into the most effective machine learning approaches for the healthcare industry, making the system relevant even for predictions of strokes.

2.4 AWS SageMaker & AWS Bedrock on Disease Prediction:

Machine learning models may be developed and deployed with outstanding flexibility using the cognitive cloud platforms AWS SageMaker and AWS Bedrock. Among other medical applications, these resources enable users to develop, fine tune, and use models for disease detection. One managed service that can assist developers in creating, fine tuning and implementing machine learning models at scale is AWS SageMaker. It supports many machine learning methods that are used for data manipulation, model training, and deployment integration in a single environment (Amazon Web Services [AWS], 2020). According to Amazon Web Services (2020), one of these built-in algorithms is called XGBoost, and it may be utilized for classification jobs such as stroke prediction. The ability of this platform to handle big data sets while

performing complex computations efficiently makes it suitable for healthcare applications where accuracy together with speed are very important.

In this investigation the authors developed a model for calculating the probability that individual patients with chronic diseases may need to be returned to the hospital has been established using SageMaker. The system demonstrated exceptional precision and provided valuable insights into the factors that lead to readmissions, making it a valuable tool for healthcare analytics. Amazon Bedrock offers a different approach by utilizing foundation models designed to provide text-based output. Consequently, it can assist in producing comprehensive reports and summaries from predictive data that are simple for medical professionals working in complicated environments. In their 2022 study, Brown et al. investigated the use of generative AI for medical report generation. When they used Bedrock to create patient reports, they incorporated clinical data and found that the statements that were produced were both logical and educational. This feature improves the interpretability of predictive models while also supporting clinical decision-making through provision of composite reports incorporating different sources of data.

2.5 Key Trends and Challenges:

The utilization of cloud platforms with cognitive computing techniques in the healthcare industry is characterized by several key trends and problems. The most notable trend is the growing use of AI and machine learning technology to enhance medical results and diagnostic accuracy. As per Srivani et al. (2023), this study involves employing deep learning techniques that have been included into (EHRs), resulting in a transformed system for providing healthcare services. But these advancements also bring with them several difficulties, notably regarding data security and privacy. Ethics issues should also be taken into consideration, according to Bhardwaj & Sharma (2020), to guarantee the confidentiality and integrity of patient data. As mentioned by Jiang et al. (2017), another challenge is the enormous expenditures required for training and infrastructure, which operate as bottlenecks to broader implementation. In addition to train their staff how to make the most of this advanced technology, hospitals must invest in high performance computing systems if they hope to see beneficial outcomes.

2.6 Identification of Research Gaps*:*

Despite the notable advancements in cloud computing and cognitive computing in healthcare, there are several research gaps. One such gap is the lack of universal application of machine learning models across different categories and environments. For instance, Beragu (2022) and ALEnezi (2019) only examined datasets, which could not accurately reflect the overall patient population. To develop models that can be used in a greater range of clinical circumstances, more work needs to be done. As Behera et al. (2019) pointed out, established protocols are also necessary for data integration and analysis. This is since their absence makes integration challenging, which impedes the successful integration of cognitive systems into healthcare services. Additionally, Bhardwaj & Sharma (2020) claim that the ethical and legal ramifications of using AI in healthcare have not been sufficiently studied or addressed. They think that since it will foster confidence between patients and providers, this should be done. Lastly, there hasn't been much research done on the integration of AI/ML with other cutting-edge technologies like blockchain and the Internet of Things. Bhuiyan & Islam, 2023. They offer a scalable cloud-based foundation for healthcare, but further research is

necessary to fully grasp the connections between these two fields and any potential obstacles to their implementation in real-world settings.

Summary of Literature Review:

There is significant promise for improving medical results, personalizing treatment programs, and increasing diagnostic accuracy through the integration of cloud platforms and cognitive computing in healthcare. The employment of deep learning techniques, the integration of electronic health records, and the growing usage of AI and machine learning technology are some of the major themes. But issues like security, data privacy, and the requirement for established protocols continue to exist. The assessment also points out areas that still require research, such as the low generalizability of models, the lack of investigation into the moral and legal ramifications, and the need for additional study on the integration of AI with cutting-edge technologies like blockchain and the Internet of Things. All things considered, cloud computing and cognitive computing have the potential to revolutionize healthcare yet to fully realize this promise, it will be necessary to solve certain obstacles and research gaps.

3. Research Methodology

This research subsection describes a methodology for using AWS SageMaker and AWS Bedrock to create, implement, and assess cognitive computing models for text generation (stroke prediction). This section offers a thorough explanation of each stage of the research in detailed procedure.

3.1 Data Collection:

The "Stroke Prediction Dataset" taken from Kaggle provided the data for this investigation. This dataset provides a wide variety of information relevant to the prediction of strokes, including both numerical and categorical variables such as age, gender, status of hypertension, heart disease, marital status, type of residence, average glucose level, (BMI), alcohol intake status, history of strokes prior to or in the family, stress levels and blood pressure measurement taken when the patient came across a healthcare professional etc. An outcome variable indicating whether someone had a stroke is also included. So, with all these provided data after undergoing critical consideration I have chosen this dataset to undergo with my research.

Ethical Declaration: The Apache License 2.0, which promotes productive and open-source programming and guarantees the supply of dependable, persistent software products, permits the usage of this dataset. It is recommended that you acknowledge the author if you use the dataset for research and that you only use it for academic use. To guarantee that the dataset was handled responsibly and in compliance with the guidelines provided by the authors, ethical considerations were taken into respect.

3.2 AWS SageMaker

The SageMaker phase consists of several steps. Initially, data is gathered and subsequently subjected to preprocessing and cleaning techniques to validate its accuracy and viability. Following that feature engineering is conducted, which could include generating interaction variables or performing other actions to enhance the quality of the dataset. Subsequently, various machine learning algorithms are compared to determine their relative performance. As shown in Fig 2, the selected model begins training until reaches its

accuracy. Once trained, it should achieve the desired accuracy threshold the model is created. After this stage, we deploy the model using Amazon SageMaker to get the endpoint. Here is the high-level overview follows,

Understanding Objectives: The key objective of this phase was to develop a prediction model that could effectively assess the risk of stroke by using many medical and social evidence. The purpose is to assist medical professionals in promptly identifying potential cases for early intervention or treatment.

Data Exploration: To understand the distribution of data and its relationships between features, EDA (Exploratory Data Analysis) was conducted. This consisting summarising the key information, highlighting any anomalies discovered, and visualising the data distributions. Using libraries like pandas and matplotlib, histograms, scatter plots & correlation matrices were created to reveal underlying patterns in the dataset.

Data Preprocessing: Data preprocessing was crucial to having a clean dataset for model. This procedure included scaling numerical characteristics to have them on a same scale, encoding categorical variables using one-hot encoding or label encoding techniques, and handling missing values by imputations. These procedures guaranteed that the standardization of data inputs into models was done consistently.

Model Training: Next development included the process of selecting a model. The XGBoost algorithm, known for its high efficiency and accuracy, has been selected for classification model training jobs. This procedure involved systematically adjusting hyperparameters to determine the optimal combination that provided the best outcomes on a validation set. SageMaker enabled this remote training capacity and automated model tuning and simplifying the handling of extensive datasets and complex models.

Evaluation Process: The evaluation process involved assessing the model's performance on a separate test dataset using various metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Confusion matrices and ROC curves were generated to visualize the model's performance and its ability to distinguish between positive and negative cases.

Deployment Phase: A model endpoint was generated after obtaining satisfactory evaluation findings. A real-time inference endpoint was developed during the deployment phase to immediately estimate stroke risks for new patient data. The implemented system was evaluated for its capacity to handle increasing workloads and maintain consistent performance across various load scenarios, utilising representative data entered by patients.

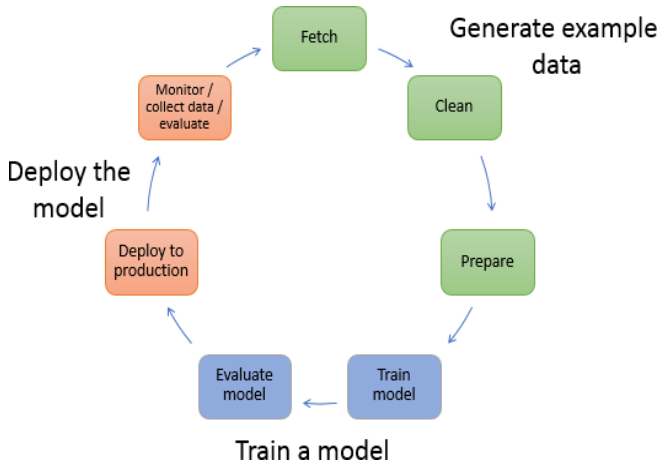

Figure 2: ML Model Workflow

3.3 AWS Bedrock:

AWS Bedrock is an AI service offered by Amazon Web Services (AWS) that allows users to utilise advanced AI models for a wide range of applications including healthcare. Bedrock simplifies the process of constructing and implementing generative AI models. The application provides a robust framework for developing complex machine learning models that can generate text that resembles human conversation using input data. Pre-trained foundation models can enhance the performance and accuracy of the model when fine-tuned for specific tasks. Its generative AI skills make it well-suited for generating relevant narrative outputs such as patient reports, summaries, and diagnostic insights. Bedrock was specifically designed to manage large-scale data processing and computing demands, making it highly suitable for extensive healthcare datasets. Bedrock seamlessly connects with other AWS services such as S3 for data storage and SageMaker for additional machine learning tasks, giving it a holistic solution for healthcare AI applications. Among the foundation models offered in AWS Bedrock, we selected a single pre-trained model. The training job was configured by establishing a provided throughput. By allocating dedicated resources to the model, we guarantee improved performance and reliability during both the training and inference stages. This model ID enables immediate engagement with the generative AI model, hence enabling the prediction of stroke risks based on real-time user input data. Furthermore, by establishing a model endpoint, Bedrock facilitates accurate and immediate evaluation of stroke risk, showcasing its potential to enhance healthcare outcomes in this regard.

3.4 Integration and Evaluation Framework:

The objective was to assess and contrast their performances in terms of precision and the time taken to make inferences for stroke risk prediction results. Streamlit was utilised as a framework for creating the user interface due to its capability to test the endpoints between SageMaker and Bedrock. This interface allows users to enter health parameters, and then utilise both SageMaker and Bedrock endpoints to obtain outputs, which can be compared in an organised way. To ensure secure connectivity with AWS services, explicit AWS credentials were established. Although not recommended for production, this method significantly accelerated the development process. The SageMaker runtime client utilised its own explicitly defined credentials while making predictions by invoking the stroke risk prediction endpoint with various inputs. Similarly, the Bedrock runtime client possesses explicit credentials that are designed to enable the creation of detailed patient reports using the generative artificial intelligence capabilities of AWS Bedrock, specifically exploiting Stroke Prediction dataset. The evaluation framework assessed the inference time of both SageMaker and Bedrock models. An essential measure is the inference time, which is determined by the duration between the submission of an input and the reception of a prediction. The comparison revealed the extent of efficiency and responsiveness in each the standard. This approach not only identified the distinguishing features of each platform but also comprehensively evaluated their performance across all dimensions inside a real-world healthcare environment.

4. Design Specifications

In this section leverages advanced machine learning techniques and cloud-based frameworks to create a stroke prediction model and risk evaluation system. The aim is to provide efficient, accurate, and scalable

predictions. This section provides a comprehensive overview of the techniques and algorithms utilised and the frameworks and tools applied, corresponding requirements and the implementation specifications.

4.1 Frameworks and Tools Used

AWS SageMaker is selected for its extensive range of machine learning functionalities, which encompass data preprocessing, model training, hyperparameter tuning, and deployment. SageMaker enables effortless interaction with a range of AWS services, guaranteeing a strong and expandable infrastructure. The primary use of AWS Bedrock is for its generative artificial intelligence (AI) capabilities, specifically in the production of comprehensive explanatory reports. Bedrock emphasise on Generative AI enables improved interpretability of prediction outcomes, which is essential for clinical decision assistance. S3 is used for data storage and AWS CloudWatch as monitoring each service. Streamlit behaves as the UI for framework, utilizing Python boto3 it facilitates the development of interactive and visually captivating dashboards for comparing models and evaluating and visualising results.

4.2 Techniques and Algorithms

4.2.1 XGBoost Algorithm in SageMaker model Training:

This project incorporates XGBoost as shown in Fig. 3, an algorithm Extreme Gradient Boosting, which is a powerful machine learning algorithm known for its efficiency and accuracy in predictive analytics. This technique attempts to reduce a loss function to optimise the predictions of the final model. XGBoost is particularly strong in its use of normalisation techniques and its ability to handle missing values. Normalisation is a technique that helps prevent overfitting by limiting the complexity of a model using L1 (Lasso) and L2 (Ridge) penalties. This leads to a model that performs effectively on data that has not been previously encountered. In addition, XGBoost could efficiently manage missing data by automatically determining the optimal course of action when encountering a missing value during the development of tree. Another notable characteristic is parallel processing, which significantly accelerates training. XGBoost efficiently utilises all CPU cores during model training, resulting in significantly quicker performance compared to classic gradient boosting approaches. SageMaker employs model training and deployments.

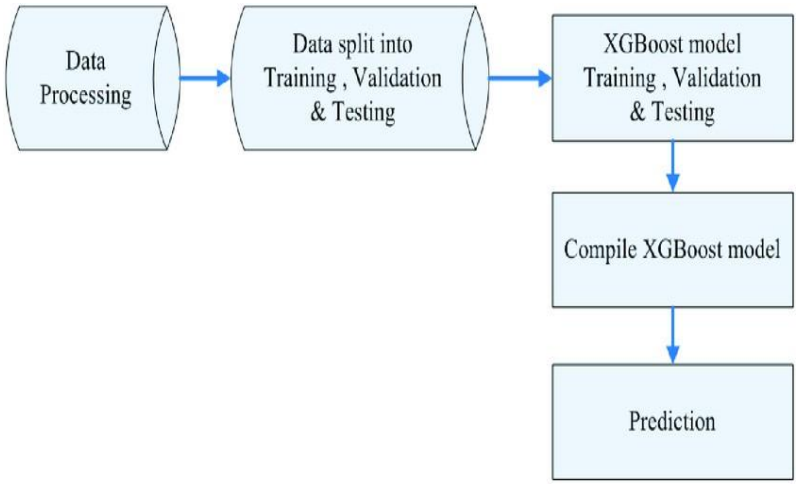

Figure 3. XGBoost Algorithm in SageMaker workflow

4.2.2 Bedrock custom model Pre-Training Set up:

The reason AWS Bedrock is utilised is because to its capacity to manage extensive generative AI jobs. We chose the Titan Text G1 Express model for its robust text generation capabilities and its ability to be customised. This model possesses the ability to comprehend and produce results that resembles human language. As a result, it is well-suited for generating comprehensive patient reports that are based on stroke prediction outcomes.

Selection of Titan Text G1 Express Model

The Titan Text G1 Express model is an advanced language model that included pre-training employing diverse text retrieved from the internet. It has demonstrated exceptional performance in several natural language processing (NLP) tasks, including as translation, summarisation among others. We selected this model because to its capacity to produce consistent and contextually appropriate texts, which are crucial for making thorough medical reports.

Continued Pre-Training

Pre-training with a stroke prediction dataset allowed us to further fine-tune the titan text g1 express model to better suit our requirements. Prior to fine-tuning the Titan Text G1 Express model & the dataset undergoes preprocessing. Essentially, we modified certain configurations of the system to enhance its comprehension and generate additional information pertaining to stroke risk prediction and healthcare terms used.

Provisioned Throughput

The system efficiently handled high volumes of requests by configuring provisioned throughput along with the pre-trained model. This involved giving a predetermined number of resources to the computer, ensuring consistent performance even during inference time when immediate responses are crucial, particularly in healthcare applications where rapid solutions are necessary. As a playground for texts entered that could be predicted based on given input prompt. Once purchased we can select our custom model to start the testing inference. Once all set, as shown in below Fig. 4, the custom model can be selected for evaluation with various test inputs

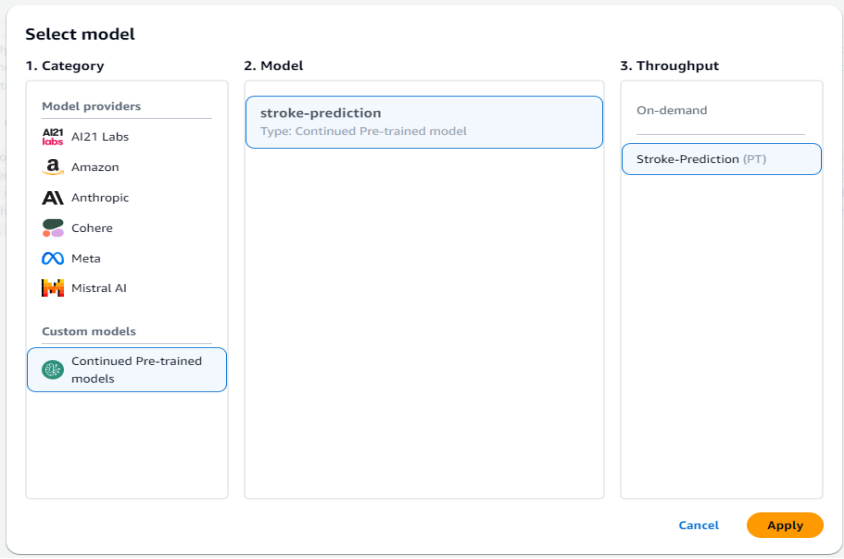

Figure 4. AWS Bedrock model workflow

4.3 System Architecture

As shown in the system architecture fig. 5, the integration of SageMaker and Bedrock for web UI comparing the performance, Streamlit application using python boto3 which enables smooth user engagement, whereby users provide input patient data obtain stroke risk predictions and generate comprehensive detailed reports. The evaluation provides a thorough presentation of comparative parameters, such as inference times and risk estimations by invoking both endpoints. By including these frameworks and tools, the design guarantees strong performance, precise prediction and secure management of healthcare data providing both efficiency and trustworthy from the application for practical use. In both AWS services S3 buckets are used according to their requirements to store the data and output. CloudWatch is utilized in both services for monitoring. This ensures that issues can be addressed and maintained in high level manner.

Data collection involves important features to prepare the data for training a model. Typically, this involves raw CSV file format that is appropriate for machine learning models, which must involve tasks like missing/ duplicate value or standardization. Using SageMaker client XGBoost model is trained, where we enhance the training process by fine-tuning hyperparameters using cross-validation techniques to attain superior accuracy. After training, the model can be deployed and accessed via a SageMaker endpoint enabling real time predictions. In Bedrock, we have done the continued pre-trained model and refined them by utilising the stroke prediction dataset. At this situation, make sure that utilizing the transformed stroke prediction dataset that was originally in csv into JSONL format. After successful model creation, this model is deployed through Bedrock's provisioned throughput to ensure efficient handling of request loads. We were developing a customised model capable of producing patient reports.

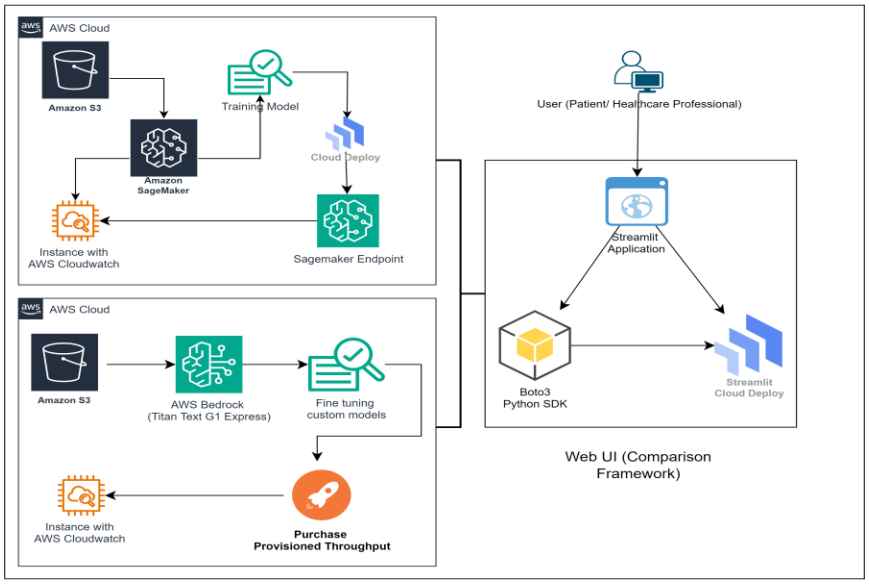

Figure 5. System Architecture Diagram

5. Implementation

The execution of the stroke prediction focused on the last steps of deploying, integrating and evaluating models to create that system capable of predicting the probability of experiencing a stroke and generating patient reports. The key components of this implementation which includes model training in AWS SageMaker, providing AI reports by invoking Bedrock and developing an interactive UI using Streamlit.

5.1 SageMaker Model Deployment

The primary achievement of this project would be bringing a model that utilized to predict stroke risk for patients. The method commenced with data preprocessing whereby the raw data was cleaned, standardised and stored in S3 for model training. The dataset acquired from Kaggle comprised various health factors such as age, gender, hypertension, and glucose levels which placed important role in predicting the risk of stroke. The capabilities of SageMaker facilitated effective data management and model training by utilising its integrated XGBoost algorithms and scalable infrastructure. After training, the model implemented on SageMaker console, allowing it to be accessed through an endpoint to invoke from outside with AWS credentials explicitly. Findings obtained throughout the training phase offered significant insights into the model's performance:

1. **Feature Importance**: The significance of feature plot identified the health parameters that had the greatest impact on stoke prediction. Understanding the decision-making process of the model and sharing insights with healthcare professionals were made easier with the use of this information.
2. **Confusion Matrix**: It offered a comprehensive analysis of model effectiveness by demonstrating elevated proportions of true positives and true negatives. This demonstrates the model endurance which accurately identifying those who are at risk of stroke.
3. **ROC Curve**: The ROC curve and its AUC score evaluate the model's high biassed ability confirming its effectiveness in distinguishing individuals who are at stroke risk or not.
4. **Precision-Recall Curve**: The utilisation of this curve proved to be extremely helpful in understanding the relationship between precision and recall facilitating selection of an ideal threshold for prediction.

5.2 Generative AI Model on AWS Bedrock

In AWS Bedrock, we have selected the Titan Text G1 Express model as primary model. The justification for choosing this one because of its stability, versatility and accessibility. However, the bedrock model should be purchased by provisioned throughput to efficiently handle large volumes of requests without any interruptions or delays. Subsequently, we implemented an adapted variant of the identical model with allocated throughput to effectively manage substantial volume of queries and guarantee prompt response times at inference phases. To optimise the performance of Titan Text G1 Express model, we conducted additional pre-training using a stroke prediction dataset. And we created custom model (Stroke-Prediction) as shown in fig. 4 with that when dealing with more complex tasks by using techniques or otherwise simulating human prompt based on given input conditions like heart attacks or food habits it became necessary for us to deploy a further among these customised models until accuracy could be guaranteed.

Model Deployment and Endpoint Configuration: Now we have customised Stroke-Prediction model implemented on AWS Bedrock. This deployment completed by configuring an endpoint that enables seamless and constant interaction with the model. The endpoint was configured to accept requests & process the input data, and provide comprehensive textual reports based on the results of stroke prediction.

5.3 Integration of Streamlit Application

To make sure secured and verified access to both SageMaker and Bedrock services, the Streamlit application explicitly configured by AWS credentials. This configuration allowed the application to invoke endpoints and handle interactions with the deployed models. The SageMaker runtime client was initialised

to communicate with the SageMaker endpoint. This client enabled instantaneous predictions by transmitting input data to the SageMaker model and obtaining predicted stroke risk score. The input data along with the stroke risk score obtained from the SageMaker model organised into a structured prompt that was appropriate for the custom Bedrock model. The prepared prompt sent to Bedrock custom model, which produced an elaborate textual report elucidating the prediction of stroke risk. The report was generated using the deployed Bedrock model and then displayed in the Streamlit application. To achieve uniform comparison, the SageMaker and Bedrock endpoints were provided with similar input data as shown in 6. By clicking Predict & Compare button the data passed to 2 models to get the results. The Streamlit application enabled seamless user interaction with the deployed models, offering a full overview of the projected along with an extensive description. The application quantified and documented the duration of each model's prediction and reporting process. The predictive performance of the SageMaker model was evaluated using conventional evaluation measures including confusion matrices, precision, recall, and F1-score. The measured metrics were shown in the Streamlit application, enabling users to directly compare the performance of the two models.

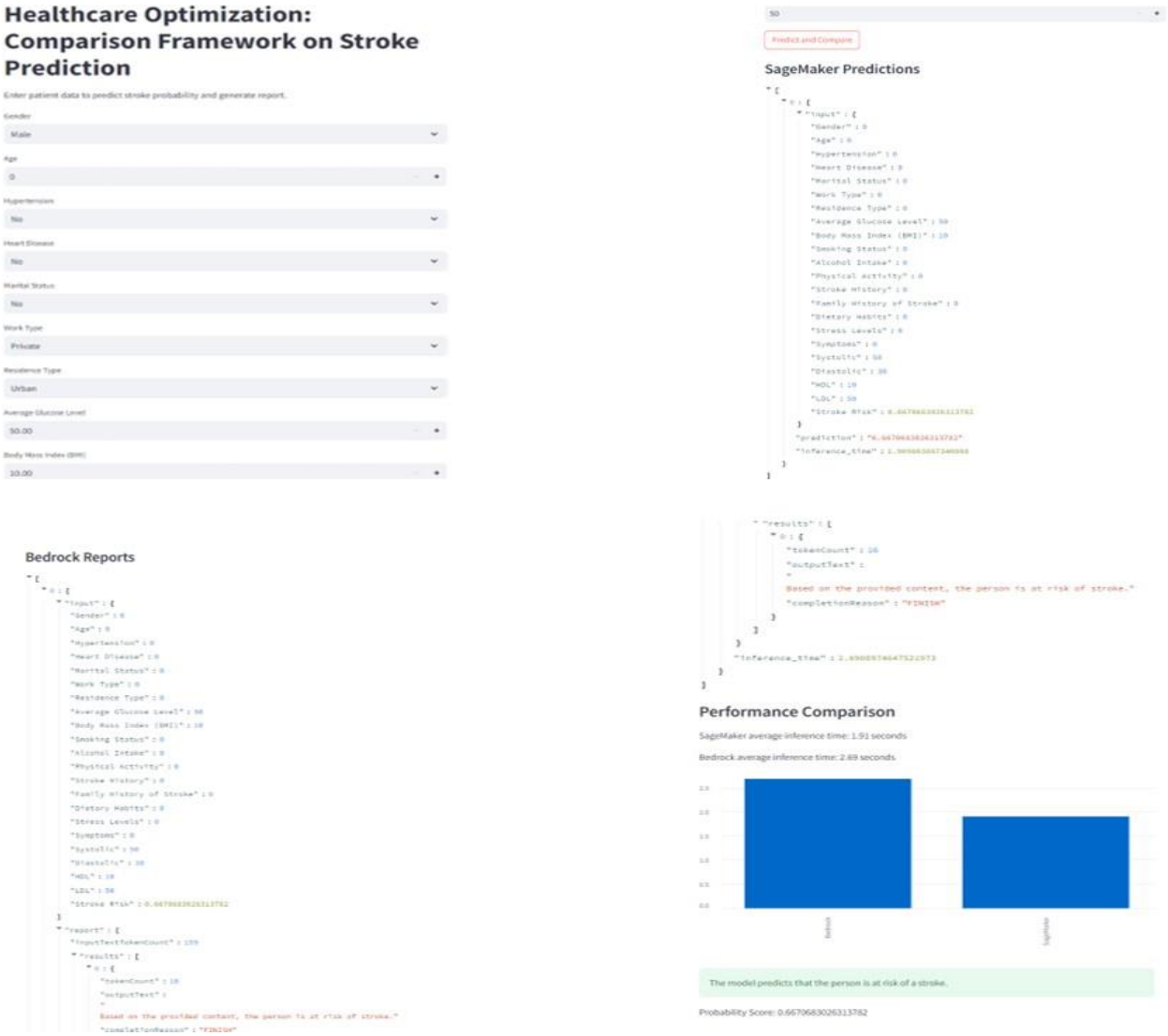

Figure 6. Streamlit UI utilized for comparison

5.4 Outputs

The key outputs of the implementation phase included the following:

Exploratory Data Analysis (EDA): Exploratory Data Analysis (EDA) allowed us to uncover valuable insights about our dataset that would have generally remained unnoticed. We also observed patterns and correlations among variables as shown in Fig 7 & 8 through visual representations. For an example I have taken BMI to illustrate the distribution of stroke risk case factors. Several significant ones include:

- Correlation Matrix: The correlation matrix highlights the relationships between several variables, providing insights into potential multiple correlations.
- BMI Distribution: Illustrates the distribution of stroke cases across different BMI levels, hence highlighting obesity as a notable risk factor.

Deployed Models: The XGBoost model deployed on SageMaker and whereas the generative AI model was hosted on Bedrock. This allows their endpoints to be applied for predictions in real time.

Interactive Application: An interactive application has been developed using Streamlit to improve user engagement and productivity. This application enables users to get input from patient, generate predictions, and explore detailed reports. As shown in Fig. 6, The results can be viewed with prediction and comparison

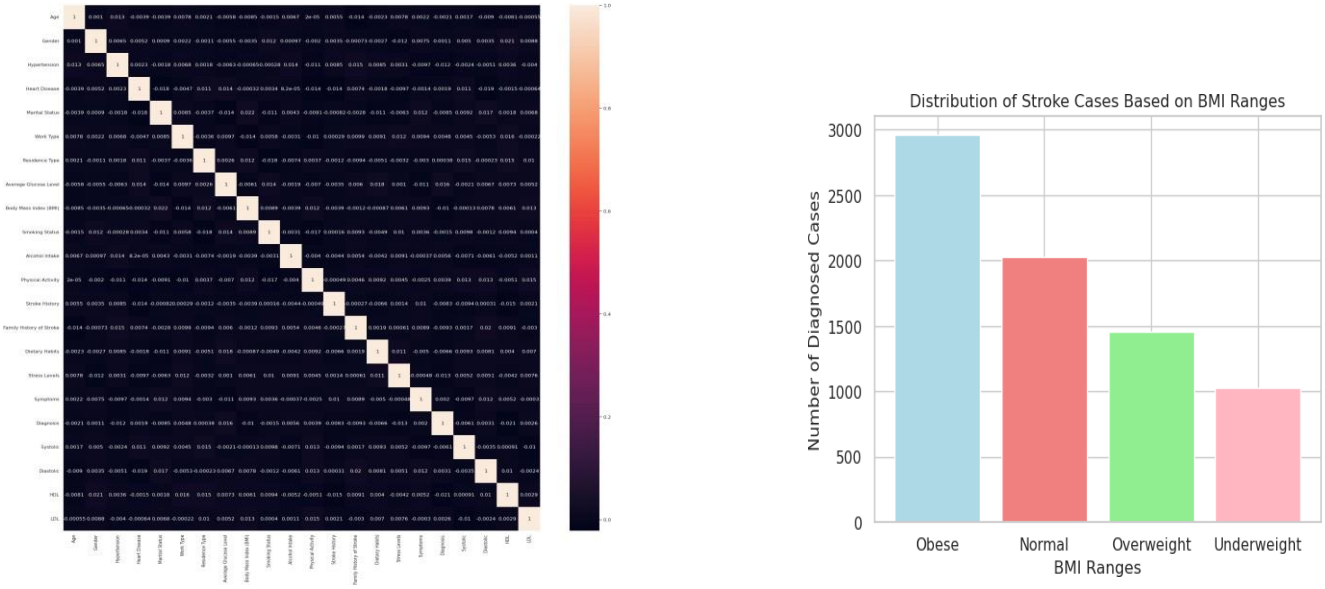

Figure 7 & 8. EDA Illustration and bar plot depicting distribution of stroke by BMI

6. Evaluation

This evaluation section is intended to give an extensive overview of the study's findings and outcomes. It measures the accuracy, efficiency and generic efficiency of the developed models in predicting the risk of having a stroke and producing comprehensive report about it. This section makes use of visualisations such as charts, graphs, and diagrams to make the study's achievements for easy understanding.

6.1 Experiment / Case Study 1: AWS SageMaker Model Evaluation

In the first case study, an assessment on AWS SageMaker stoke prediction (Text Generation) is evaluated. In this phase, an alternative set of variables from the patients is used for testing to verify the model's accuracy and other performance metrics.

Confusion Matrix: This illustration in fig. 9 shows how accurately people were categorised as either at risk or not at risk of stroke. It displays true positives, true negatives, false positives, and false negatives providing a measure of the model's recall, precision, and F1-score.
ROC Curve: The ROC curve shown in fig. 10 illustrates how effectively this model can distinguish between positive and negative classes by plotting sensitivity data. Using this method, AUC curve is also computed to show overall performance.
Performance Metrics: Several remarkable efficiency metrics were recorded, including recall rate, accuracy rate, precision rate and overall f1 score among others. The results from SageMaker inference time and its effectiveness in predicting stroke risk are highlighted.

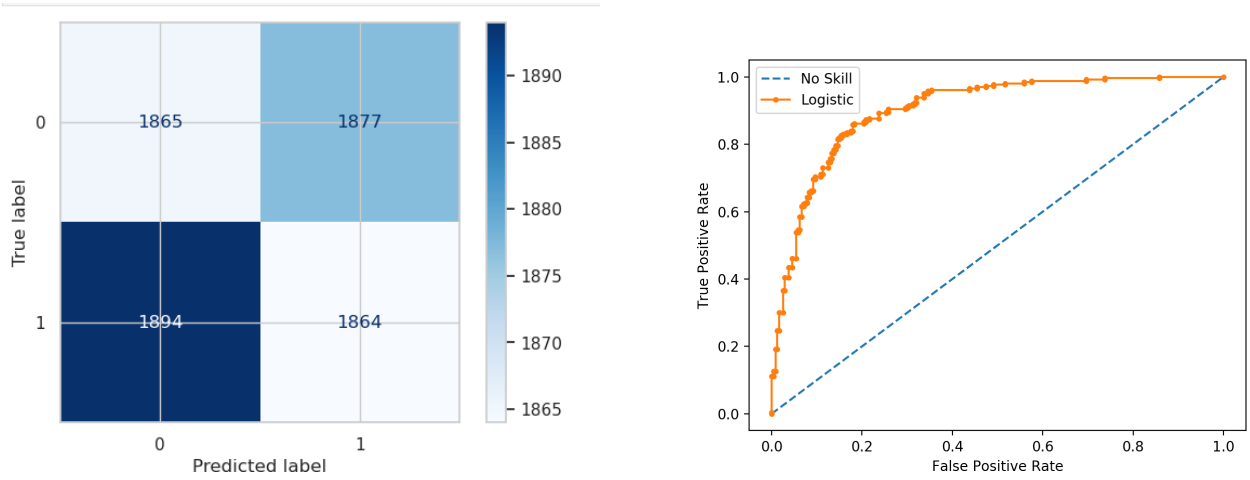

Figure 9 & 10. Confusion Matrix Graph & ROC Curve

6.2 Experiment / Case Study 2: AWS Bedrock Evaluation

The second case study evaluates the performance of the generative AI model on AWS Bedrock specifically the Titan Text G1 Express model, which is a unique pre-trained model called Stroke - Prediction. As shown in fig. 11, this custom model provides comprehensive textual reports based on its prediction results verified.

Textual Report Generation: The Bedrock model is evaluated by using the identical patient data is utilised for the SageMaker model. The resulting reports are examined for consistency, contextual relevance, and transparency. It is determined whether the model can explain the patient's risk variables in an understandable way to healthcare professionals.
Performance Comparison: The average time required for the Bedrock model to make inferences is compared to the SageMaker model. Furthermore, the assessment of the precision of the generated reports in terms of accurate diagnosis the risk of stroke is performed.

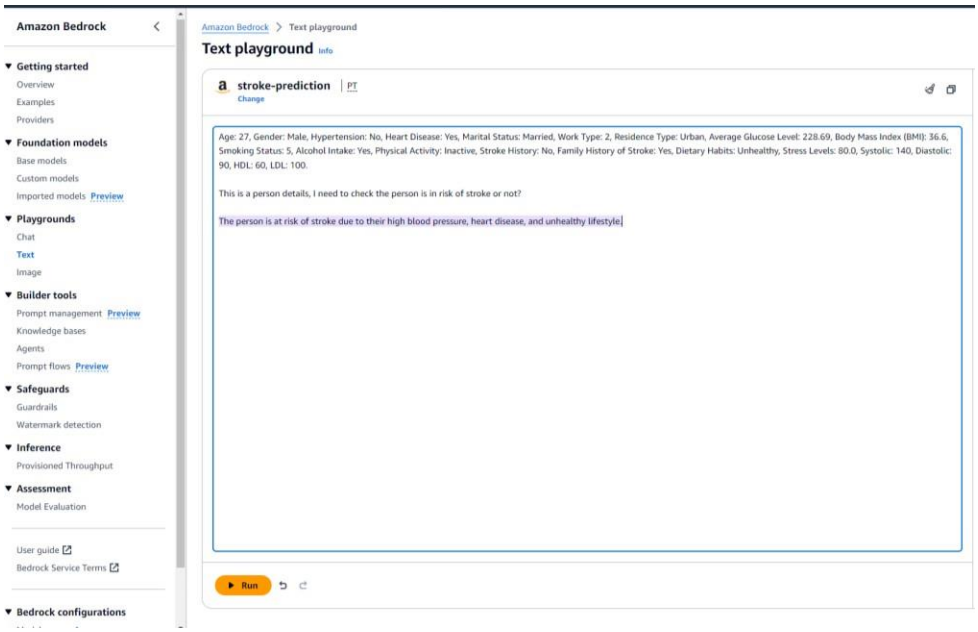

Figure 10. Custom model Evaluation in AWS console

6.3 Discussion

The SageMaker model achieved exceptional accuracy and efficiency in predicting stroke risks, as indicated by its confusion matrix, ROC curve, and performance measurements. Conversely, the Bedrock model has been producing significant outcomes that are relevant to healthcare practitioners. In terms of inference times, SageMaker exhibits a marginal speed benefit, although at an insignificant scale. Although both platforms are well-suited for real-time applications. The models are integrated into Streamlit application to enable usage by healthcare providers who need predictions and reports based on patient data input. The development of this project placed significant emphasis on ethical aspects such as safeguarding patient data privacy and security, as well as addressing and minimising bias in AI prediction. Making sure the models were transparent and understandable was paramount to assist people understand the reasoning behind healthcare companies decision-making processes. Although preprocessing data and fine-tuning the model present challenges, our research supports previous studies that show an impressive enhancement in diagnostic accuracy when cognitive computing systems are combined with generative AI models, while adherence to ethical principles.

7. Conclusion and Future work

The objective of this study must develop and evaluate a test generation (stroke prediction) comparison system utilising AWS SageMaker & AWS Bedrock. The primary objective was to enhance the prediction of diagnosis using machine learning and generative AI models while providing healthcare providers with comprehensive without difficulty of understandable data. The project successfully accomplished these objectives by implementing two reliable models. The primary findings indicate that beneficial in their respective contexts. Metrics like as precision, recall, and F1-score give empirical evidence supporting the excellent accuracy achieved by the SageMaker model in predicting the risk of stroke. In addition to its ability

to provide precise and appropriate descriptions of patient's risk factors, the Bedrock model also stands out for its exceptional coherence in the development of reports, far exceeding any other known system. While the former is marginally faster than the latter during inference time, both are suitable for real-time applications that require speed. Although the difference between them may not appear significant, the integration of Streamlit has enabled the development of a user-friendly interface, making it clear how useful these tools can be. This is particularly beneficial for healthcare providers who require quick access.

Although the study was successful, it did have several limitations, such as the difficulty of preprocessing big datasets and the necessity of fine tuning the models. To overcome these constraints, future research could focus on enlarging the training dataset and investigating advanced methods to enhance the performance of the model. Incorporating image processing capabilities in AWS Bedrock would enhance in means of diagnosis by addressing an extended range of diseases, which would explore it more inclusive. Subsequent investigations could concentrate bringing together of image processing and text generation to develop a comprehensive healthcare solution. It would include managing more intricate data kinds and could necessitate advanced GPU capabilities and optimised Amazon cost control. Investigating the economic stability of this integrated system has the potential to bring about major improvements in healthcare technology, offering healthcare professionals with vast techniques for diagnosing.

References

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., & Al Yami, M. S. (2023). Revolutionizing healthcare: The role of artificial intelligence in clinical practice. BMC Medical Education, 23(1), p. 689.

Behera, R. K., Bala, P. K., & Dhir, A. (2019). The emerging role of cognitive computing in healthcare: a systematic literature review. International Journal of Medical Informatics, 129, pp. 154-166.

Beragu, S. (2022). Effective use of Cloud Computing and Machine Learning Technologies for Smart Healthcare Applications. Dublin: National College of Ireland.

Brown, T., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, pp. 1877-1901.

Build generative AI applications on Amazon Bedrock — the secure, compliant, and responsible foundation | Amazon Web Services. (2024, June 29). Amazon Web Services. Available at: https://aws.amazon.com/blogs/machine-learning/build-generative-ai-applications-on-amazon-bedrock-the-secure-compliant-and-responsible-foundation/.

Castaneda, C., Nalley, K., Mannion, C., Bhattacharyya, P., Blake, P., Pecora, A., ... & Suh, K. S. (2021). Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine. Journal of clinical bioinformatics, 5, pp. 1-16.

Custom models - Amazon Bedrock. (n.d.). Available at: https://docs.aws.amazon.com/bedrock/latest/userguide/custom-models.html.

Davenport, T. and Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), pp. 94-98.

Dash, B. (2024). Zero-Trust Architecture (ZTA): Designing an AI-Powered Cloud Security Framework for LLMs' Black Box Problems. Current Trends in Engineering Science, 4, p. 1058.

Huang, X., et al. (2022). Novel insights on establishing machine learning-based stroke prediction models among hypertensive adults. Frontiers in Cardiovascular Medicine, 9, p. 901240.

Islam, M. M., & Bhuiyan, Z. A. (2023). An Integrated Scalable Framework for Cloud and IoT-Based Green Healthcare System. IEEE Access, 11, pp. 22266-22282.

Jiang, F., et al. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4).

Kamnitsas, K., et al. (2017). Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis, 36, pp. 61-78.

Mohajeri, M.A. (2024). Leveraging large language model for enhanced business analytics on AWS.

Ramraj, S., Uzir, N., Sunil, R. and Banerjee, S., 2016. Experimenting XGBoost algorithm for prediction and classification of different datasets. International Journal of Control Theory and Applications, 9(40), pp.651-662.

Sharma, R., & Ghosh, U. B. (2022). Cognitive computing driven healthcare: A precise study. In: Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis. Singapore: Springer Nature Singapore, pp. 259-279.

Srivani, M., Murugappan, A., & Mala, T. (2023). Cognitive computing technological trends and future research directions in healthcare—A systematic literature review. Artificial Intelligence in Medicine, 138, p. 102513.

What is Amazon SageMaker? - Amazon SageMaker. (n.d.). Available at: https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html

Wolf, P. A., D'Agostino, R. B., Belanger, A. J., & Kannel, W. B. (1991). Probability of stroke: a risk profile from the Framingham Study. Stroke, 22(3), pp. 312-318.