ASSIGNMENT 01 DAMG7245 - Big Data System and Intelligent Analytics

[Team Members](#_a45ovnkp3i1n)

[Press Release](#_yiowlgfmjv4o)

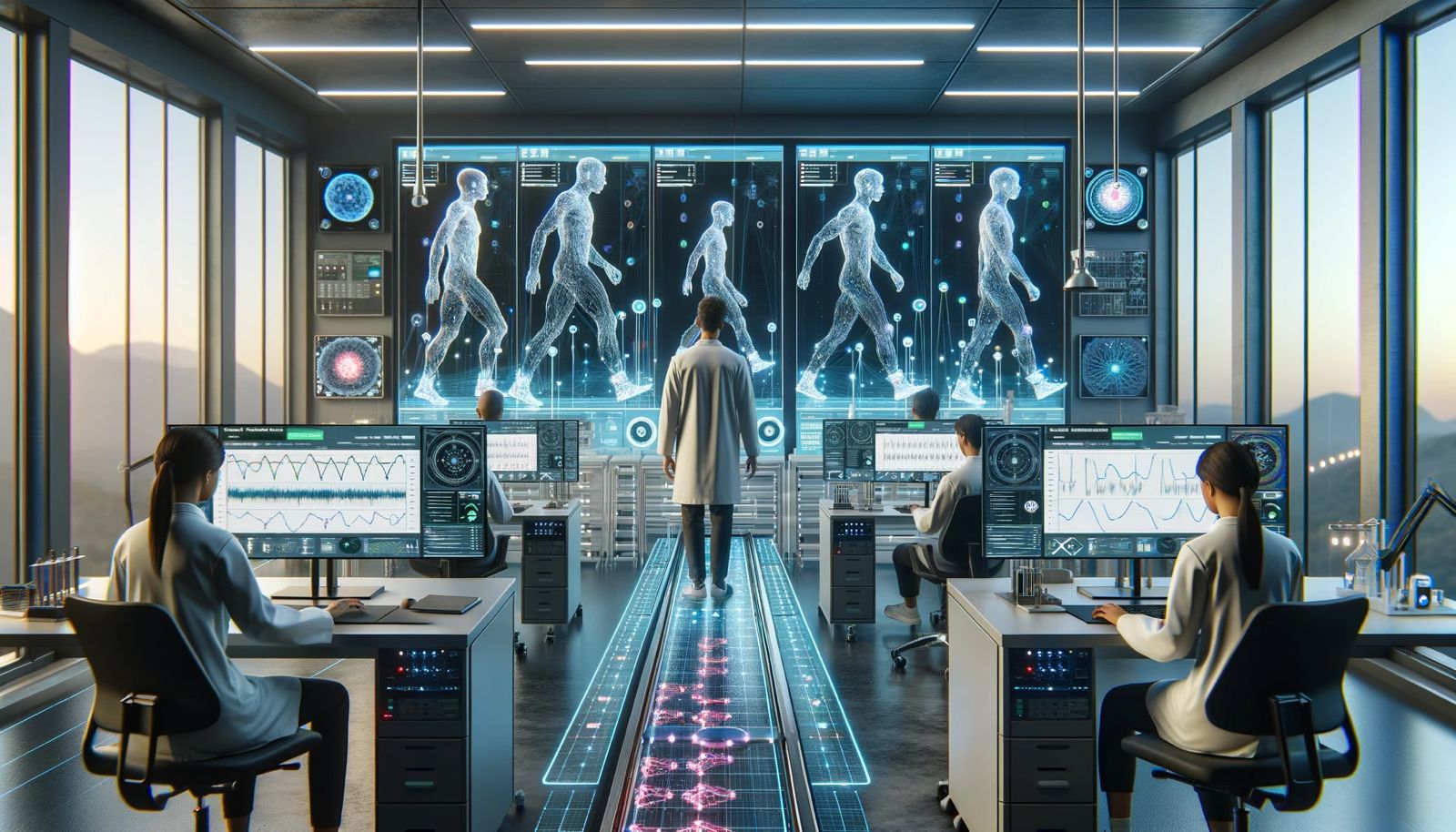
[An External FAQ](#_9b6597m8s7c6)

[An Internal FAQ](#_p7yro9cdkgid)

[Architectural Diagram](#_rvs4u9iyecld)

# Case Study 1

## Revolutionizing Early Detection



### Team Members

#### 

| Name | NUID |
| --- | --- |
| Nidhi Kulkarni | 2827172 |
| Riya Singh | 2826938 |
| Deepakraja Rajendran | 2845156 |

# 

# 

# 

# Press Release

## **Seeing What Was Unseen: Early Detection Reimagines Neurodiversity**

Empowering Healthcare Professionals: A Revolutionary Tool Simplifies Autism Detection and Diagnosis in Neurodivergent Individuals

The rapidly evolving landscape of neurodevelopmental disorders calls for innovative solutions, particularly in the case of two major players, autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD), both experiencing a concerning surge in prevalence.

At the forefront of innovation, we proudly introduce a revolutionary diagnostic tool meticulously designed for the precise identification of physiological malfunctions, with a specific focus on autism. This groundbreaking solution surpasses the accuracy of conventional diagnostic practices, empowering healthcare professionals in the early detection and diagnosis of neurodivergent disorders.

### 

### **The Challenge:**

Autism: Recent statistics from the CDC indicate that 1 in 36 children now carries an ASD diagnosis, with a disproportionately higher rate of 4 times in boys. This surge cuts across racial, ethnic, and socioeconomic boundaries, emphasizing the urgent need for effective detection and intervention.

Comorbidity with ADHD: Compounding the challenge, up to 70% of individuals diagnosed with ASD also struggle with ADHD, creating a complex diagnostic landscape that requires sophisticated tools for accurate identification and treatment.

### 

### **The Solution:**

Early Detection is Crucial: Extensive research highlights the early years as a "benign stage" for intervention in neurodivergent disorders, offering a window of opportunity for better outcomes and empowering individuals with timely support.

Our advanced diagnostic tool, powered by state-of-the-art machine learning, operates on the principle of collecting gait data from patients. This data is then meticulously analyzed, providing healthcare professionals with unparalleled insights into the early stages of neurodivergent disorders, particularly autism.

### 

### 

### **Conclusion:**

As we navigate the challenges posed by the rising prevalence of neurodivergent disorders, our commitment to early detection and intervention remains unwavering. This revolutionary tool aims to redefine the standards of diagnosis, providing healthcare professionals with a powerful ally in the quest for improved outcomes and enhanced support for individuals navigating the complexities of neurodiversity.

# External FAQ

**Q1: How does the advanced diagnostic tool work?**

A: The advanced diagnostic tool operates by using a well-trained model. It collects data in a clinical lab setting, predicts results using artificial intelligence, and then undergoes analysis by experts. This analysis eventually guides the determination of suitable therapies or treatments based on the generated report.

**Q2: What makes this tool more advanced than existing diagnostic practices?**

A: Unlike traditional diagnostic methods that primarily rely on visual analysis and often miss early signs of issues, our advanced diagnostic tool stands out. Our AI application is designed to detect even subtle neurological signs, making it a unique and groundbreaking innovation in the diagnostic market. This approach is entirely new to the medical industry, offering a fresh and highly effective perspective for identifying and addressing health concerns, especially in their early stages.

**Q3: What evidence supports the claim of greater precision than conventional practices?**

A: To back up our assertion of superior precision compared to traditional methods, we plan to incorporate the latest research accuracy rates into our application's model. This involves staying abreast of the most recent developments in the field and leveraging the latest information to improve the accuracy and efficacy of our application. By consistently integrating the most recent findings, we can confidently state that our approach outperforms conventional practices, ensuring precise and dependable results.

**Q4: How will the doc collect the data?**

A: To gather the necessary data, a dedicated laboratory is established within the hospital. Trained laboratory staff then work directly with patients to collect the required information. Once this data is obtained, it is input into the application. The application, leveraging its capabilities, generates a comprehensive diagnostic report. This report is then forwarded to the healthcare professional for thorough analysis. Based on the findings, the doctor can recommend and initiate the appropriate course of treatment tailored to the individual patient's needs. This process ensures a thorough and personalized approach to healthcare.

**Q5: How long will the process take and when will I get the report?**

A: The duration of the process relies on the cooperation level of each patient. Typically, it takes around 45 to 75 minutes to gather the necessary data. After that, you can expect to receive your diagnostic report within 2 to 3 days, depending on the hospital's availability.

**Q6: What are the potential risks associated with the adoption of this tool?**

A: Identified risks include initial skepticism from healthcare professionals, data privacy concerns, and potential technical glitches. We have comprehensive risk mitigation strategies in place, and ongoing monitoring will address emerging issues.

**Q7: What is the feedback mechanism from healthcare professionals, and how are we adapting to it?**

A: We actively seek feedback through regular surveys, forums, and direct communication with healthcare professionals. This iterative feedback loop informs our updates and improvements to meet the evolving needs of the medical community.

**Q8: Is the diagnostic tool suitable for all age groups, or is it specifically designed for children?**

A: While our initial focus is on early detection in children, the diagnostic tool is adaptable and applicable to a broader age range. The machine learning model is designed to interpret gait data across different developmental stages, ensuring versatility in its application.

**Q9: Can the diagnostic tool be utilized in telehealth settings, or is it limited to in-person assessments?**

A: While the tool is initially designed for in-person assessments in clinical settings, we recognize the potential for telehealth applications. Our team is exploring ways to adapt the tool for remote use, considering the evolving landscape of healthcare delivery

# Internal FAQ

**Q1: How scalable is the infrastructure supporting the diagnostic tool?**

A: The infrastructure is designed with scalability in mind, allowing for increased data processing demands as the tool gains wider adoption. Regular assessments and upgrades are part of our ongoing strategy.

**Q2: How can healthcare professionals integrate this tool into their practices?**

A: We are open to collaboration and discussions with healthcare professionals, providing comprehensive evidence supporting the precision of our diagnostic tool. Training sessions and support will be offered for seamless integration into existing practices.

**Q3: Can patients access the diagnostic tool directly, or is it only available through healthcare professionals?**

A: The tool is intended for use by trained healthcare professionals in a clinical lab setting. Direct patient access is currently not part of the scope, as the expertise of professionals is essential for accurate data collection and interpretation.

**Q4: How can healthcare institutions integrate this tool into their existing systems?**

A: We provide comprehensive training sessions and support for the seamless integration of the diagnostic tool into existing healthcare systems. Our technical team collaborates with institutions to ensure compatibility and efficiency.

**Q5: What privacy measures are in place to protect patient data during the diagnostic process?**

A: We prioritize patient privacy and comply with all relevant regulations. Data collected is handled securely, and access is restricted to authorized healthcare professionals. Our commitment to data security is unwavering.

**Q6: What is the workflow for the product?**

A: The solution architecture is designed to leverage a sophisticated machine-learning model for predicting or classifying autism in patients. This model is deployed on AWS SageMaker, ensuring robust performance and scalability. The core components of the architecture and their interactions are outlined as follows:

1. Streamlit Web Application:

The user-friendly web application is developed using Streamlit, known for its capacity to swiftly create interactive apps. This application serves as the primary interface for patients or healthcare providers to input patient data.

2. Data Collection and Transmission:

Patient data, cleaned waveform data from sensors monitoring patient mobility, is securely collected through the Streamlit application. This data is then directly transmitted to the AWS SageMaker endpoint, where the machine-learning model is hosted.

3. Machine Learning Model on AWS SageMaker:

AWS SageMaker hosts the predictive model, meticulously trained to classify or predict autism based on the input data. SageMaker's robust and scalable environment ensures that the model's performance is optimal and that it can handle varying loads, making real-time predictions with high accuracy.

4. Report Generation and Visualization:

Once the prediction is made by the model, the result is sent back to the Streamlit application. A comprehensive report is dynamically generated within the app, presenting the prediction and relevant insights in an easily interpretable format. Users can view the report directly on the application.

5. Email Dispatch via AWS Simple Email Service (SES):

For enhanced accessibility, the report can also be emailed to the patient or concerned parties. This is seamlessly handled by integrating the AWS Simple Email Service (SES) with the Streamlit app, ensuring secure and reliable delivery of the reports.

6. Authentication and Security:

Security is paramount, especially considering the sensitivity of health-related data. A proxy service - NGINX- is configured in front of the Streamlit application to manage authentication. It interfaces with AWS Cognito to authenticate users, fetch tokens, and ensure that only authorized individuals can access or input data, maintaining the integrity and confidentiality of the patient information.

7. Data Storage on AWS S3:

All patient records, including newly input data and generated reports, are securely stored in AWS S3 buckets. S3 provides a highly durable and available storage solution, ensuring that the data is safely stored and can be accessed or retrieved when needed.

8. Monitoring and Logging with AWS CloudWatch:

To maintain the reliability of the production environment and ensure smooth operation, AWS CloudWatch is employed. It continuously monitors the system, running diagnostic tests, and logging processes. This not only helps in keeping track of the system's performance but also aids in quick troubleshooting in case of any anomalies or issues.

The architecture is meticulously designed to ensure a seamless, secure, and efficient workflow, from data input through the web application to the machine learning prediction on AWS SageMaker, and finally to the communication of the results via the report in the app and through email, all while maintaining high standards of security and monitoring.

**Q7: Why Streamlit?**

A: Streamlit integrates seamlessly with data science and machine learning workflows, making it easy to visualize data, interact with models, and share results. The learning curve for it is smaller and it is ideal for data analysis. It is hosted on AWS EC2 service.

**Q8: What does the application look like?**

A: The application is designed to streamline the workflow for lab technicians, simplifying the data entry process for predictive analysis. The user-friendly interface requires the input of essential patient details, including the Patient's Name, Age, Contact Information, and specific Mobility Statistics. After the necessary data is inputted, technicians can initiate the prediction process with a simple click on the 'Classify' button.

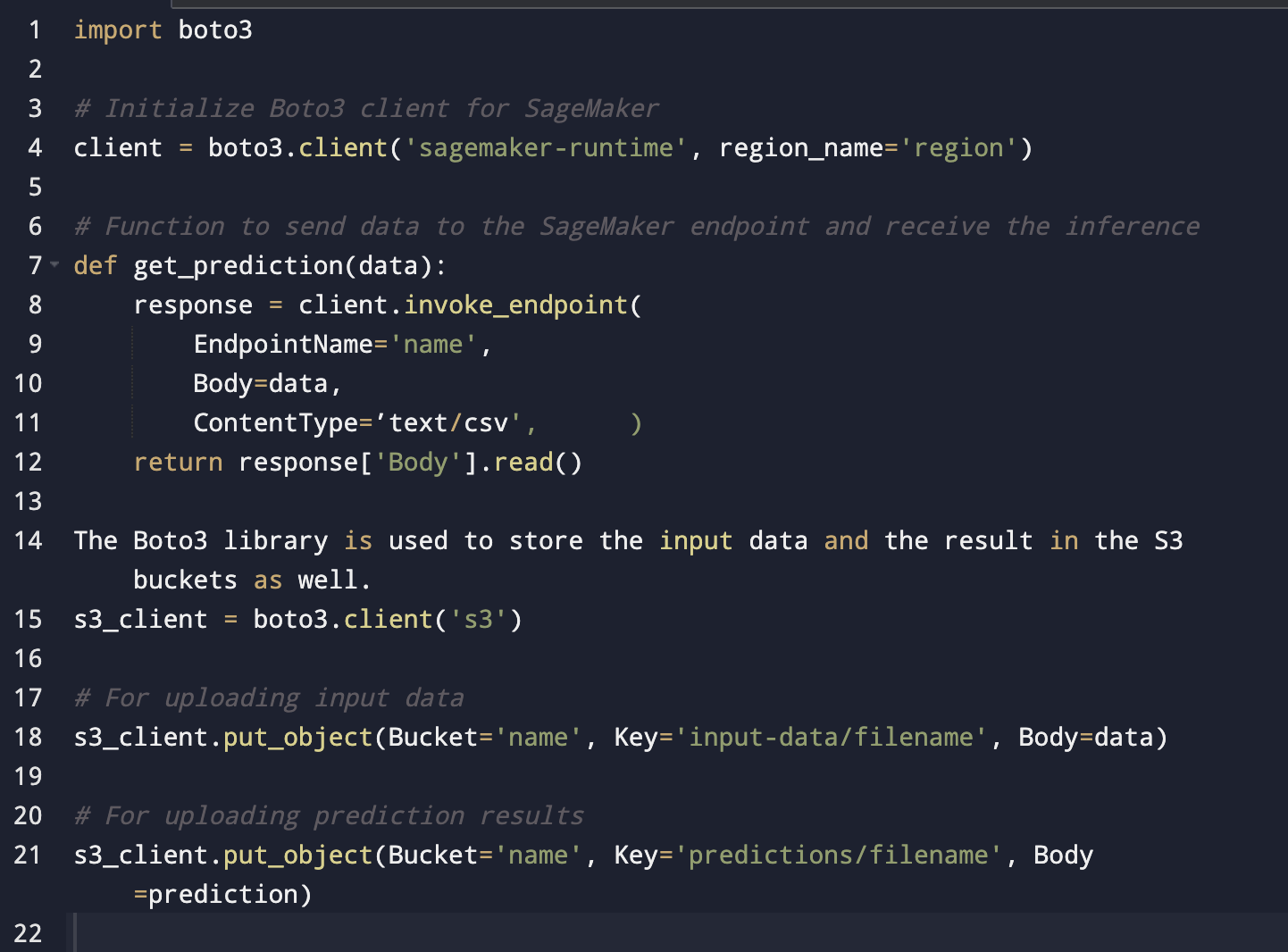
Upon completion of the predictive analysis, the application generates a comprehensive report. This report, which is immediately available for download, offers detailed insights, including the classification results and the accuracy percentage of the prediction. The report can be generated using libraries like ReportLab, FPDF, or WeasyPrint. For enhanced convenience and patient engagement, the report is also automatically dispatched to the patient's email address through the AWS Simple Email Service, ensuring a seamless and efficient communication flow.

The design of the application emphasizes not only operational efficiency and user-friendliness for lab technicians but also a commitment to providing clear, actionable information to patients, fostering transparency and trust in the diagnostic process.

**Q9: How is data collected and transmitted from the source?**

A: Once the user inputs the data, it is aggregated in the form of CSV for the Sagemaker model. Boto3, the AWS SDK is used to invoke the endpoints (SageMaker/S3) and transmit the data.

The real-time data is sent to SageMaker to invoke the inferences.



**Q10: How is the machine learning model selected and used for the prediction?**

A: Model selection will be done based on the unique characteristics of the dataset and its features and the performance of each model during training and validation processes. During training and validation, various models are evaluated, considering factors like accuracy and precision. With diverse datasets from individual customers like banks, the selection process is tailored to specific requirements. Thorough research precedes deployment, ensuring the chosen model aligns precisely with the dataset's nuances and the objectives of the prediction task.

**Q11: How does authentication using a proxy server work?**

A: The authentication proxy [Nginx](https://www.nginx.com/) is used in front of your Streamlit app. The proxy can handle authentication with AWS Cognito and then forward the traffic to the Streamlit app. When the user accesses the application, the proxy intercepts the request and redirects the user to AWS Cognito for authentication. After successful authentication, AWS Cognito redirects the user back to the proxy with an ID token. The proxy validates the ID token and, if valid, forwards the request to the Streamlit app. This method abstracts the authentication process away from the Streamlit app, which can be beneficial for keeping the app simple and focused on its primary functionality.

**Q12: Why NGINX as a proxy server?**

A: NGINX can handle a large number of concurrent connections with a low memory footprint, making it efficient. NGINX’s modular architecture allows it to be extended with third-party modules, adding functionalities such as additional security layers, authentication mechanisms, or even modifying or inspecting requests and responses. It can be hosted on AWS EC2.

**Q13: How will the process be monitored?**

A: A monitoring and validation framework is established within AWS CloudWatch, tailored to ensure the reliability and accuracy of the deployed machine learning model. This sophisticated framework is programmed to autonomously execute a series of diagnostic test cases every half hour, serving as a consistent and methodical evaluation mechanism for the model's predictive performance.

The test suit comprises two types:

Positive Classification Test Case: This case meticulously examines the model's aptitude to correctly identify instances indicative of autism. Success is predicated on the congruence between the model's prediction and the expected outcome, thereby validating the model's sensitivity and diagnostic precision.

Negative Classification Test Case: This case rigorously assesses the model's ability to accurately discern non-autistic instances. This is crucial for ensuring the model's specificity and minimizing the incidence of false-positive results.

With each test case execution, AWS CloudWatch meticulously orchestrates the process, capturing extensive data including

Operational Metrics: Vital performance indicators from the production environment, such as latency, throughput, and error rates, are systematically logged. This data offers invaluable real-time insights into the operational health and efficiency of the system.

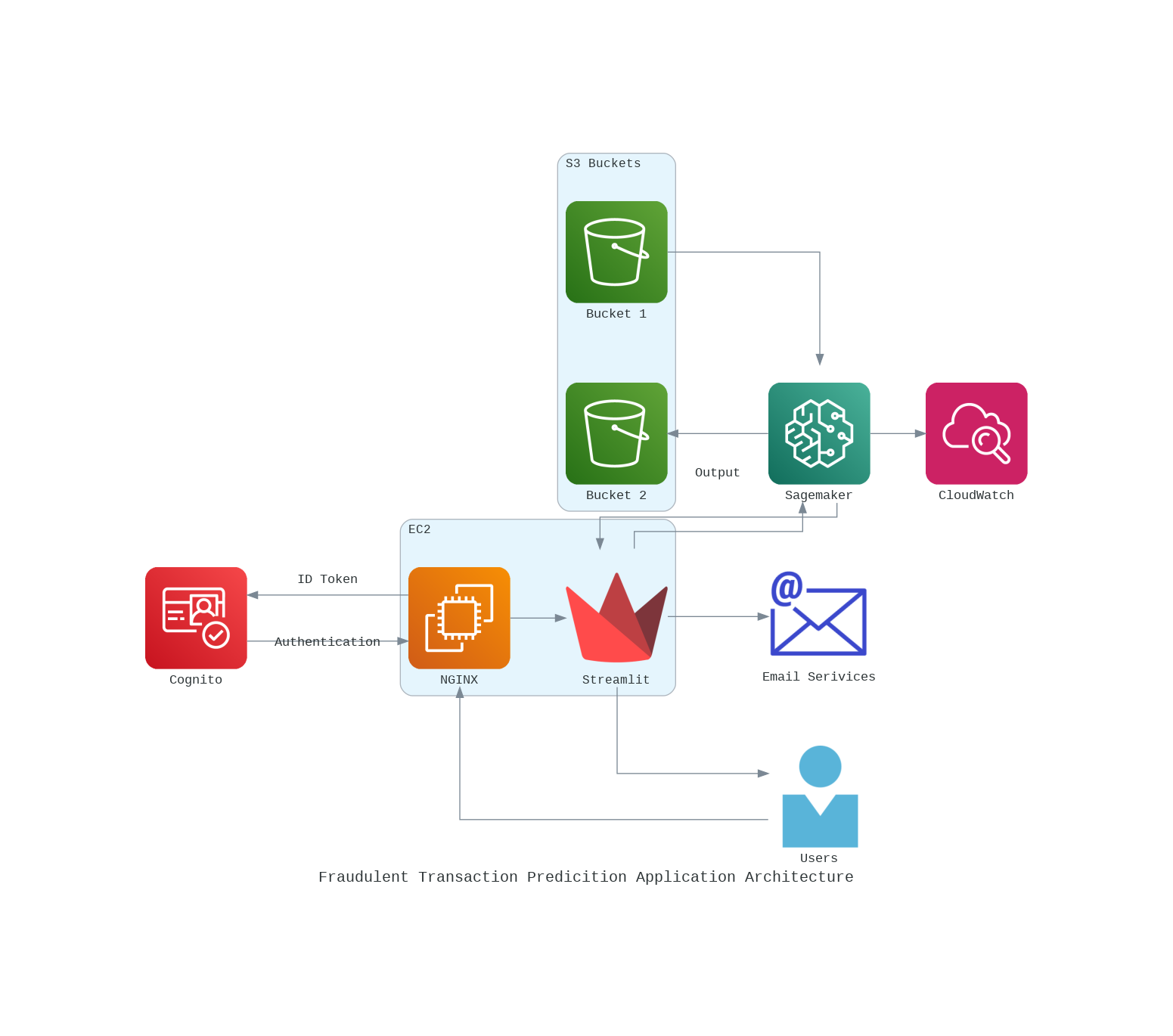
Test Execution Logs: Comprehensive logs detailing each test execution are methodically maintained. These logs include the input data, the model's predictive output, the expected result, and the final verdict of the test (pass/fail). This constitutes a thorough record for quality assurance, operational auditing, and compliance.

Anomaly Detection and Alerting Mechanisms: The system is adeptly configured to identify anomalies or test failures, triggering immediate alerts. This feature is instrumental in ensuring rapid response and remediation, maintaining the integrity and performance of the model in the production setting.

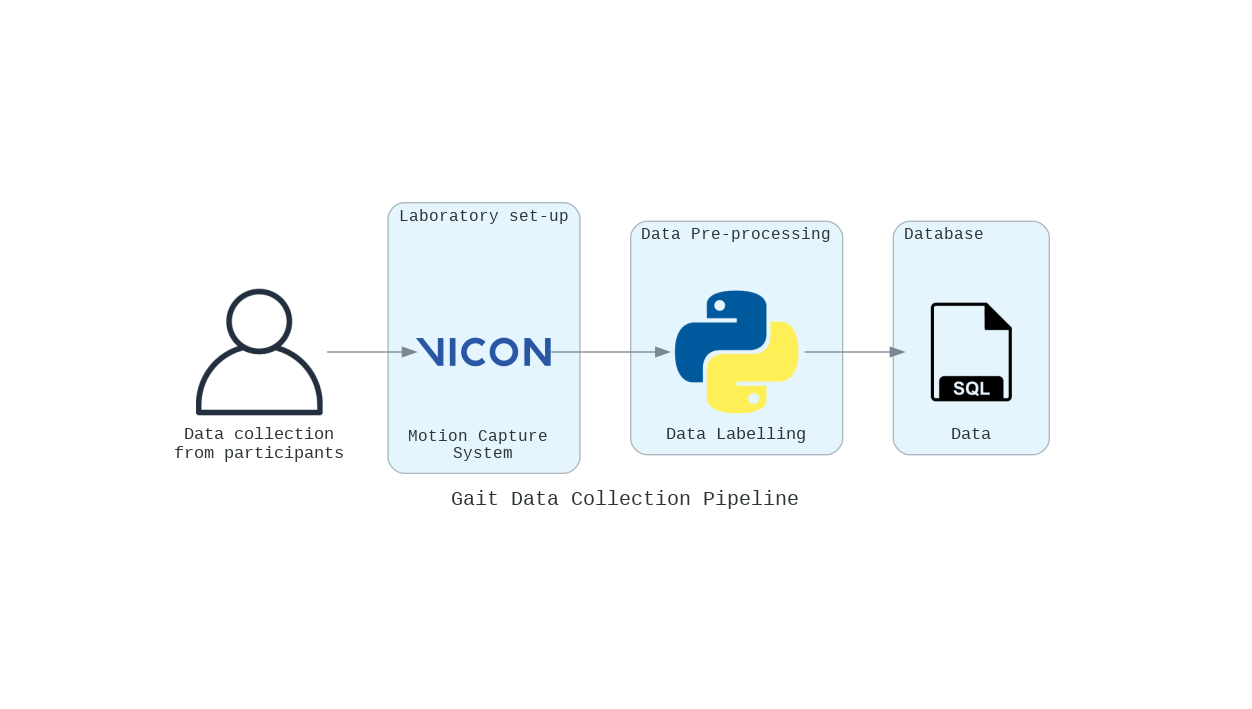
**Q14: How’s the data collection process done?**

A: The data collection process involves inviting volunteering participants to undergo a meticulously controlled series of laboratory steps, capturing gait data using Vicon motion capture systems. These systems record body movements, transforming the acquired waveform data into numerical form. Subsequently, the data undergoes expert-led preprocessing and labeling within the laboratory. This comprehensive approach ensures the dataset is primed for machine learning model training. The labeled data, containing gait information, becomes the foundation for a model capable of accurately interpreting and predicting patterns in human movement dynamics.

# Architectural Diagram



## Data Pipeline:



### 

### 

### 

### 

### 

### 