Project Proposal
Big Data Pipeline for Multi-Source Breast Cancer Risk Analysis
Group 5

Group Members:
Basanth Periyapatna Roopa Kumar (018202600), Manav Rajesh Anandani (018194917), Mayank Kapadia (017687x735), Nischitha Nagendran (018174104)

1. Abstract

Breast cancer is still a leading cause of death, highlighting the crucial need for scalable and rapid data analytics in healthcare. This project provides a strong big data architecture that uses Hadoop Distributed File System (HDFS) for scalable data storage, Apache Hadoop and Spark for fast batch processing, and Apache Kafka combined with Spark Streaming for real-time analytics. Our approach solves the substantial issues of integrating and analyzing structured, semi-structured, and unstructured healthcare datasets at large volumes and speeds. Our system effectively manages heterogeneous healthcare datasets by building a consistent, scalable pipeline for simultaneous batch and real-time data processing, resulting in continuous, actionable insights into breast cancer risk factors. The resulting system exhibits higher performance, fault tolerance, and scalability, emphasizing big data technology's transformative potential to greatly improve preventative healthcare skills and decision-making processes.

2. Motivation
Breast cancer's persistent worldwide health burden reveals major gaps in current healthcare analytics, owing mostly to the limits of traditional data management methods. Traditional technologies are unsuitable for the scale, complexity, and real-time nature of today's healthcare data, restricting the ability to intervene on time. Motivated by these constraints, our research creates a ground-breaking big data pipeline specifically built to handle heterogeneous, high-volume medical datasets efficiently. Using cutting-edge distributed technologies, our solution tackles current inadequacies by providing rapid data integration, real-time responsiveness, and resilient scalability. The proposed pipeline intends to greatly advance healthcare analytics by enabling physicians to make informed, proactive decisions based on timely insights gleaned from large-scale multi-source data.

3. Literature Survey
[1] J. J. Jeong et al.: The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.5M Screening and Diagnostic Mammograms.
This research study describes a large-scale (3.4 million pictures from 116,000 patients) ethnically heterogeneous breast imaging dataset with thorough lesion-level annotations and pathologic results. This dataset addresses the limitations of existing, often homogeneous and smaller datasets by ensuring equitable representation of African American and White women, who are frequently underrepresented while having poorer breast cancer outcomes. EMBED comprises over 40,000 annotated lesions with areas of interest (ROIs) linked to structured imaging descriptors, as well as 56 ground truth pathologic outcomes, which are critical for training and testing deep learning models for breast cancer screening. The curation process entailed dealing with diverse image kinds (2D, DBT), extracting ROIs from screen save images, and overcoming problems associated with DICOM metadata variability and image normalization among manufacturers. Finally, EMBED hopes to assist the creation of more generalizable and egalitarian AI models for breast cancer diagnosis, and a portion of the dataset is publicly available for research.
Link: https://pubmed.ncbi.nlm.nih.gov/36721407/

[2] S. Zhou et al.: CancerBERT: A Cancer Domain-Specific Language Model for Extracting Breast Cancer Phenotypes from Electronic Health Records.
CancerBERT is a specific language model designed to accurately extract essential breast cancer information, known as phenotypes, from Electronic Health Records (EHRs). The researchers developed this model by training it on a significant amount of text data, primarily from breast cancer patients. A large portion of their effort involves improving the model's capacity to grasp medical language by incorporating bespoke vocabulary containing cancer-specific phrases identified through both expert knowledge and the frequency with which these terms appeared in medical texts. CancerBERT outperformed other general and biological language models in recognizing critical characteristics including hormone receptor status and tumor size. Notably, the versions of CancerBERT that

included the customized vocabularies performed better, demonstrating the value of a language model that is uniquely adapted to the medical domain.
Link: https://pubmed.ncbi.nlm.nih.gov/35333345/

[3] C. Tai, H. Gunraj, and A. Wong: Cancer-Net BCa-S: Breast Cancer Grade Prediction Using Volumetric Deep Radiomic Features from Synthetic Correlated Diffusion Imaging.
The authors introduce Cancer-Net BCa-S, a volumetric deep radiomics framework designed to non-invasively predict breast cancer severity using synthetic correlated diffusion imaging (CDI). The traditional Scarff-Bloom-Richardson (SBR) grading system typically involves invasive biopsies, increasing patient discomfort and healthcare expenses. This study solves that constraint by using advanced image analytics to reliably detect cancer grades in the absence of biopsies. When compared to traditional MRI approaches, Cancer-Net BCa-S has a much greater prediction accuracy (87.7%), outperforming existing MRI-based standards. The method entails synthesizing diffusion signals, extracting volumetric deep radiomic characteristics, and using a neural network-based grade prediction model. This study not only improves non-invasive diagnostic skills, but also illustrates the revolutionary power of big data and volumetric imaging approaches in clinical breast cancer grading.
Link: https://arxiv.org/abs/2304.05899

4. Methodology

4.1 Data Collection & Preprocessing:

We continue to look for appropriate datasets for deployment at this stage. We will finalize the dataset when we proceed. However, we plan to incorporate:

- Medical Imaging Data (e.g., MRIs, mammograms) from public repositories such as CBIS-DDSM or synthetic data.
- Electronic Health Records (EHR) for patient history and medical history, perhaps obtained from MIMIC-III or other publicly available medical data.
- Wearable Sensor Data (e.g., sleep, physical activity, heart rate) for real-time analysis. Synthetic logs will be generated if real data is unavailable.

Preprocessing Plan:

- Convert structured data (EHR) to Parquet format for efficient querying.
- Normalise and segment unstructured data (images).
- Save wearable sensor data in Kafka topics for real-time analytics.

4.2 Scalable Storage Architecture:

We will implement a distributed storage system to cater to the diverse and large datasets.

- HDFS (Hadoop Distributed File System): For batch processing EHR and image data sets.
- Cloud Storage (S3/GCP Storage): For high-scalability access to medical imaging data.

4.3 Big Data Processing:

We plan to employ a hybrid model that blends batch and real-time processing:

Batch Processing (Hadoop MapReduce)

- Process past patient records to derive long-term risk trends.
- Extract meaningful features such as age, tumor size, hormone levels, genetic factors.

In-Memory Processing (Apache Spark)

- Perform feature engineering and train machine learning models on Spark MLlib.
- Use classification models such as Random Forest, XGBoost.

4.4 Real-Time Streaming Integration:

Real-time monitoring is a significant part of the project, though the real-time streaming source has yet to be determined. The strategy includes:

- Apache Kafka as the real-time data ingestion platform for wearable data (heart rate, sleep monitoring, etc.).
- Spark Streaming to process the incoming data and detect anomalies in real time.

If real wearable data is not accessible, we will utilize simulated sensor readings to assess the pipeline.

4.5 Machine Learning & Deep Learning

- ML Models (EHR + Wearables): Random Forest, XGBoost for risk prediction.
- Deep Learning (Medical Images): CNN-based models (e.g., ResNet, EfficientNet).

Explainable AI (XAI):

To ensure trustworthiness and transparency in predictions:

- SHAP (SHapley Additive Explanations) will be used to explain patient risk scores.
- LIME (Local Interpretable Model-Agnostic Explanations) will interpret predictions for tabular EHR data.
- Grad-CAM (Gradient-weighted Class Activation Mapping) will highlight important regions in medical images for doctor interpretation.

5. Deliverables and Milestones

Milestone	Deliverables
Week 1 and Week 2	● Dataset exploration and finalization (medical imaging, EHR, wearable sensors, and synthetic data). ● Set up HDFS and cloud-based storage systems (such as AWS S3). ● Initial preprocessing: convert EHR to Parquet, standardize and segment pictures, and configure Kafka topics for streaming data.
Week 3 and Week 4	● Create a batch processing pipeline with Hadoop MapReduce to examine historical patient data for trend extraction. ● Feature engineering with Apache Spark. ● Initial data pipeline testing and validation to ensure correctness and scalability.
Week 5 and Week 6	● Deployment of a real-time streaming infrastructure (using Apache Kafka in conjunction with Spark Streaming). ● Validate streaming data ingestion with wearable or simulated sensor data. ● Run the initial anomaly detection and real-time analytics tests.
Week 7 and Week 8	● Combine batch and real-time analytics into a unified pipeline. ● Use and validate ML classification algorithms (Random Forest, XGBoost) and CNN models (ResNet, EfficientNet). ● Use explainability techniques (SHAP, LIME, Grad-CAM) to enhance model transparency.
Week 9	● Final testing and performance improvement for the combined pipeline. ● Complete the project documentation and technical report. ● A project presentation showcasing pipeline capabilities and analytical insights.

6. Team Members and their Roles

Team Member	Roles
Basanth Periyapatna Roopa Kumar	● Data collection and preprocessing the data ● Setting up Kafka for real-time data ingestion
Manav Rajesh Anandani	● Training & Developing ML and DL models ● Implement explainable AI
Mayank Kapadia	● Big Data Processing (Spark) ● Perform Feature Engineering

Nischitha Nagendran	● Implementing HDFS and storage architecture ● Develop Batch Processing

7. Relevance to the course

Big Data management of quantity, speed and type supports all essential concepts within the "Big Data Technologies and Applications" course. Here in this project, we are covering all the essential V's of Big data, making this more relevant to the course. Real-time processing along with batch modes implemented through Hadoop and Apache Spark technologies enable the project to develop a data pipeline that conducts early breast cancer identification using medical information. The system controls real-time data processing with Apache Kafka while meeting multiple execution requirements using structured, unstructured, and semi-structured information.
The implementation of big data processing technologies into wearable sensor data, medical images, and EHRs exactly follows the fundamental concepts taught in the course. This project shows how Big Data solutions tackle real-world issues by detecting breast cancer early and fulfills the key goal of utilizing Big Data in meaningful case scenarios.

8. Rubric Criteria Fulfillment

- Abstract: Clearly defines the project's purpose, scope, and technologies, emphasizing big data pipeline integration for breast cancer risk assessment.

- Motivation: Emphasizes the importance of scalable and real-time big data analytics in healthcare, aligning with course subjects.

- Literature Survey: Includes IEEE-formatted research articles that demonstrate a thorough comprehension of current big data approaches and datasets for breast cancer diagnosis and analytics.

- Methodology: Experimental design is explicitly documented, including big data technologies (Hadoop, Spark, Kafka), specific data processing methodologies (batch and streaming analytics), and stated assessment methods.

- Deliverables: Outlines milestones for each project step, such as data preprocessing, system installation, pipeline integration, testing, and technical paper submission.

- Team Members and their Roles: Ensures uniform workload distribution among all team members, with each actively involved in critical big data project components, including storage, batch processing, streaming integration, and system validation.

- Relevant to the course: The project is strongly related to big data principles covered in the curriculum, such as distributed storage (HDFS), scalable processing (Hadoop, Spark), real-time analytics (Kafka, Spark Streaming), and pipeline orchestration.

- Technical Difficulty: Reflects high technological complexity by combining numerous sophisticated big data technologies into a unified, scalable pipeline with real-time capabilities, displaying ambitious but achievable goals.

- Novelty: Innovative combination of large data batch and streaming approaches, including adaptive pretreatment and explainability, distinguishes from standard fragmented analytics methods.

- Impact: High potential for practical healthcare applications, including improved breast cancer risk assessment and early diagnosis through advanced big data analytics, with great publication potential.

- Heilmeier Catechism: Effectively resolves each point of the Heilmeier Catechism, offering precise definitions of project scope, feasibility, risks, expenses, and evaluation indicators.

9. Technical Difficulty

While working on this project, the following technical difficulties can arise:

- **Working with integrating data:** In this project we will be functioning with structured and unstructured data from diverse medical sources (which includes wearable data, medical imaging and EHRs). We require some synchronization methods and rules to carefully manage this data.
- **Working with highly scalable data stores:** We will be storing our data on HDFS and cloud systems which would be quite challenging while dealing with this much amount of big data.
- **Performing data preprocessing:** Feature extraction and model training can be difficult and challenging because of noise or inconsistent or missing data which may pertain in medical pictures and wearable sensor data.
- **Real time Analytics challenge:** Since we will be detailing with high-velocity data streams our system must have exact latency and performance measures when we are working on real time analytics using kafka and spark streaming.
- **Maintenance of accuracy and explainability:** Maintenance of accuracy for medical data while guaranteeing the explanation using SHAP, Grad-CAm can turn out to be difficult.

10. Novelty

This initiative distinguishes itself through an original synthesis of big data frameworks designed specifically to promote real-time breast cancer analysis. Unlike previous approaches that were limited to discrete batch analytics, our solution delivers a unified and scalable pipeline that seamlessly blends large-scale historical analytics with continuous, real-time anomaly detection. The design uses adaptive preprocessing techniques, distributed storage, and streaming technologies to efficiently manage varied healthcare data sources. Furthermore, by incorporating explainability directly into the analytics pipeline, the initiative improves clinical interpretability, allowing for actionable and transparent medical insights. Finally, our initiative proposes a highly practical yet innovative paradigm for rethinking preventative healthcare through scalable, fast, and interpretable big data analysis.

11. Impact

This project advances healthcare analytics by providing a scalable big data pipeline that combines hybrid batch and streaming analytics, allowing for precise and rapid breast cancer risk assessment. The system effectively deploys Hadoop, Spark, and Kafka, demonstrating practical solutions to high-volume data processing challenges in clinical settings. Furthermore, the use of explainable analytics in this architecture boosts clinical trust by providing transparent, interpretable insights. Finally, the approach establishes a strong foundation for data-driven preventive healthcare, directly improving patient outcomes through timely intervention.

12. Heilmeier Catechism

1. What are you trying to do?

We intend to create a scalable and integrated big data analytics pipeline that uses distributed processing technologies, notably Hadoop, Apache Spark, and Kafka, to enable rapid and exact evaluation of breast cancer risk using large-scale healthcare data analysis.

2. How is it done today, and what are the limits of current practice?

Currently, healthcare businesses rely on fragmented data analytics workflows with limited real-time capabilities and scalability. Traditional analytical frameworks struggle to handle large, diverse medical information, resulting in delayed diagnosis and limited proactive interventions.

3. What is new in your approach and why do you think it will be successful?

Our strategy combines batch processing and real-time streaming analytics, using Hadoop for historical data management, Spark for high-performance calculations, and Kafka for real-time data input and analysis. The revolutionary combination of these technologies allows for efficient large-scale data handling and continuous monitoring, resulting in greatly improved early detection and risk assessment.

4. Who cares? If you are successful, what difference will it make?
Healthcare providers, hospitals, and clinical decision-makers are all heavily invested in this breakthrough. Successful adoption will result in improved patient outcomes due to timely intervention, lower healthcare costs, and increased clinical efficiency because of real-time, data-driven decision support.

5. What are the risks?
Key challenges include successfully gathering and integrating three distinct datasets—medical imaging, electronic health records, and wearable sensor data—all of which raise concerns about data quality, privacy, and interoperability. aWe may encounter difficulty in establishing uniform data formats, handling incomplete or missing information, and meeting healthcare-specific privacy standards. Furthermore, ensuring real-time scalability and compliance with healthcare data rules may present additional practical challenges during the project's execution.

6. How much will it cost?
To keep costs reasonable, we want to mostly use free-tier cloud resources, such as AWS or Google Cloud Platform trial services, open-source technologies (Hadoop, Spark, Kafka), and publicly available data. While this drastically lowers direct costs, we may only incur minimal expenditures if we exceed our free-tier quotas for cloud storage or computation resources. Once the specific scope and infrastructure requirements have been identified, a thorough budget estimate will become evident.

7. How long will it take?
The big data pipeline's development and deployment will take about 8 to 10 weeks total. This include data gathering, preparation, architectural integration, performance assessment, and final deployment.

8. What are the midterm and final “exams” to check for success?
Midterm milestones include successfully creating the data pipeline architecture (storage, batch processing) and completing the first real-time data integration tests. The final success criteria include establishing efficient real-time analytics, high system scalability and responsiveness, and proven accuracy in identifying breast cancer risk markers from large-scale data.

13. References

[1] J. J. Jeong et al., "The EMory BrEast imaging Dataset (EMBED): A Racially Diverse, Granular Dataset of 3.5M Screening and Diagnostic Mammograms," Scientific Data, vol. 10, Article 132, 2023. doi: 10.1038/s41597-023-02020-6.

[2] C. Tai, H. Gunraj, and A. Wong, "Cancer-Net BCa-S: Breast Cancer Grade Prediction using Volumetric Deep Radiomic Features from Synthetic Correlated Diffusion Imaging," arXiv preprint arXiv:2304.05899, 2023.

[3] S. Zhou et al., "CancerBERT: A Cancer Domain-Specific Language Model for Extracting Breast Cancer Phenotypes from Electronic Health Records," arXiv preprint arXiv:2108.11303, 2021.

[4] F. J. Alonso, M. Belgiu, and J. Kang, "Leveraging Apache Spark for Real-Time Data Analytics in Healthcare Systems," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 480-492, 2021. doi: 10.1109/BigData50022.2020.9377948.

[5] A. Ghosh and D. Dasgupta, "Big Data Analytics in Healthcare: Opportunities and Challenges," in Proc. IEEE Int. Conf. on Big Data (Big Data), Atlanta, GA, 2020, pp. 2177–2185. doi: 10.1109/BigData50022.2020.9377948.