National College of Ireland

Machine Learning & PBFT Blockchain Methodology on AWS for Proteomics Analytics

MSc Research Project
Cloud Computing

Sravanthi Challa
Student ID: 21156239

School of Computing
National College of Ireland

Supervisor: Shaguna Gupta

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sravanthi Challa
Student ID:	21156239
Programme:	Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Shaguna Gupta
Submission Due Date:	14/12/2023
Project Title:	Machine Learning & PBFT Blockchain Methodology on AWS for Proteomics Analytics
Word Count:	5926
Page Count:	18

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:	Sravanthi Challa
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
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.	☐

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):	

Machine Learning & PBFT Blockchain Methodology on AWS for Proteomics Analytics

Sravanthi Challa
21156239

Abstract

Proteomics research, in particularly on PROTEIN IDENTIFICATION, an area that is expanding very Fastly and has the potential to completely change how the world think about biology and medicine. In order to help proteomics analytics to reach new high level, this research project looks into the very helpful technology combination of Machine Learning (ML), Practical Byzantine Fault Tolerance (PBFT) Blockchain, and Amazon Web Services (AWS). This research focuses on how we use machine learning (ML) algorithms like KNN (K-Nearest Neighbors), Neural Networks, Decision Trees, Random Forest, Logistic Regression, Random Forest or Logistic Regression can interpret complex patterns in proteomics datasets so that it can improve the accuracy of protein identification and quantification. Also, this study includes about how to take in the PBFT Blockchain into the proteomics data management system to work on AWS cloud to obtain latency. Through confidentiality and Encryption, this integration works to strengthen data integrity, security, and traceability throughout the proteomics data process, resulting in increased dependability and credibility in research findings. An interaction between PBFT Blockchain's strong security features and ML-driven precision in protein analytics is the main expected results, which gives us more precise protein identification and quantification while guaranteeing unmatched levels of data integrity and security in proteomics research.

Keywords- Proteomics Analytics, Machine Learning, PBFT Blockchain, AWS.

1 Introduction

Proteomics - study of the large set of different structures and functions of proteins, that supports many cellular processes makes proteomics an important field for biological systems. As mentioned by Osbourn (2014) and Chen, Wang, Li, Xiao, Gao, Huang, Zhao, Wu, Xu, Chen and Li (2022) proteins are vital indicators that power a range of biological functions in living organisms. According to Kumar et al. (2022) there are still a lot of issues with protein identification and quantification methods that need to be resolved in order for proteomics research to improve. To resolve these issues, this project works at how well machine learning (ML) approaches may improve the accuracy of protein identification and quantification in proteomics data. Proteomics data management is enabling data integrity, security, and provenance by integrating blockchain technology in response to the increasing need for safe and dependable data handling methods Chen, Zhang, Liu et al. (2022)Liu et al. (2023). As such, the central research questions focus on how

machine learning techniques can be applied to improve the precision of protein identification and quantification, as well as how blockchain technology can be incorporated to strengthen proteomics data management.

1.1 Research Question

How can machine learning techniques be effectively applied to proteomics data to improve protein identification and quantification accuracy? How can blockchain technology be integrated into proteomics data management to ensure data integrity, security, and provenance?

1.2 Problem Statement

Problem: Proteomics, a data-intensive field, struggles with accurate protein identification and data security. Traditional methods lack precision and scalability, risking data integrity and undermining scientific trust. This research employs machine learning for precise protein analysis and integrates blockchain technology to fortify data integrity and security.

Relevance: This research addresses critical issues in proteomics, essential for advancing our understanding of complex biological systems. Applying machine learning enhances protein identification, benefiting medicine, biotechnology, and genetics. Integrating blockchain technology secures data, ensuring trust and reproducibility. This interdisciplinary approach can transform proteomics and influence various scientific and medical domains.

1.3 Proposal / Solution

Solution: Our project, " Machine Learning & PBFT Blockchain Methodology on AWS for Proteomics Analytics," presents a holistic solution to the data-intensive challenges in proteomics research. We will collect, clean, and integrate diverse proteomics data, employ Machine Learning for improved analysis, integrate Blockchain for data authenticity, and utilize AWS for scalable computational power.

Relevance: This research addresses critical issues in proteomics, essential for advancing our understanding of complex biological systems. Applying machine learning enhances protein identification, benefiting medicine, biotechnology, and genetics. Integrating blockchain technology secures data, ensuring trust and reproducibility. This interdisciplinary approach can transform proteomics and influence various scientific and medical domains. With the goal of revolutionising the precision, dependability, and security of protein identification and data management, this work makes a substantial contribution to the scientific literature by fusing cutting-edge technologies—blockchain and machine learning—with proteomics research.

2 Background of Protein Data

The field of proteomics stands as a cornerstone for studying the complex mechanisms underlying cellular functions and biological processes. Proteins, as the workhorses of biological systems, play diverse roles in cellular structure, function, signaling, and regulation. Understanding the complexity of these molecular machinery requires an in-depth

knowledge of proteomics, which is the study of all the proteins that are present in a cell, tissue, or organism.

2.1 Protein

Proteins exhibit diverse levels of internal structure and organization (Figure 1), primarily contains of amino acids as their fundamental building blocks. These amino acids, form connections through peptide bonds, constituting the primary structure of a protein. The functional configuration of a protein involves intricate three-dimensional architecture, characterized by -helices and -sheets—recognized as secondary structure elements. Subsequently, the ultimate three-dimensional layout of an individual protein denotes its tertiary structure. Conversely, an amalgamation of multiple proteins forms a quaternary structure. While predicting the three-dimensional structure from the amino acid sequence is only feasible for smaller proteins, larger structures need experimental methods Neumann (2014).

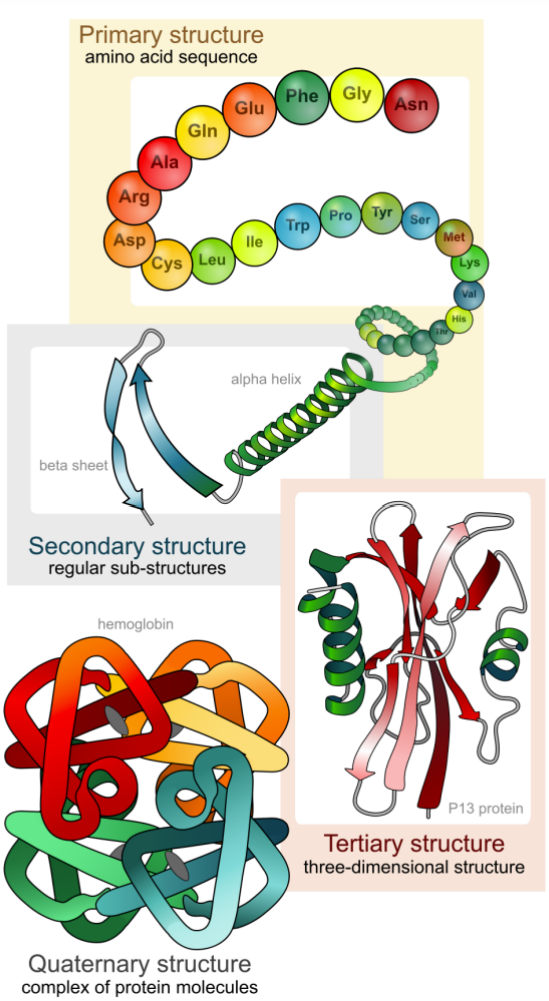

Figure 1: Stages of Protien Structure (Taken from Wikipedia[3])

Techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy drive experimental structure education, providing static snapshots of stable conformations. Actually, the importance of protein structures earned its recognition

through the 1962 Nobel Prize, awarded to John Kendrew and Max Perutz for their pioneering work in unveiling the first atomic structure of a protein via X-ray crystallography Neumann (2014).

2.2 PDB – Protein Data Bank

The Protein Data Bank (PDB) is a comprehensive archive of protein structures that are determined through various experimental and computational methods. PDB files are created by researchers who have determined the three-dimensional structure of a protein and are then deposited in the PDB. The process of creating a PDB file typically involves the following steps:

2.2.1 Structure Determination

Scientists can use a variety of methods to determine the three-dimensional structure of a protein, including:

- X-ray crystallography: In this method, the protein is crystallized and X-rays are shot at it. The diffraction patterns are then analyzed to determine the positions of the atoms in the protein's structure.
- Nuclear Magnetic Resonance Spectroscopy (NMR): This method utilizes the magnetic properties of protons in amino acids to determine the positions of the atoms in the protein's structure.
- Cryo-electron microscopy (cryo-EM): This method involves freezing a sample of the protein and then imaging it with electrons. The images are then used to determine the positions of the atoms in the protein's structure.

2.2.2 Structure Validation and Annotation

Once the protein structure has been determined, it is validated for accuracy and completeness. The structure is then annotated with additional information, such as the organism from which the protein was isolated, the method used to determine the structure, and the experimental conditions used.

2.2.3 Submission and Review

The annotated protein structure is then submitted to the PDB. The PDB has a rigorous review process that ensures that all submitted structures meet high standards of quality and accuracy.

2.2.4 Public Release

Once the structure has been reviewed and approved, it is released to the public as a PDB file. PDB files are freely accessible to researchers and other interested parties.

PDB files are essential tools for researchers studying proteins. They provide valuable information about the structure and function of proteins, which can be used to understand their role in various biological processes, develop new drugs and therapies, and design

new materials. This research project is also using PDB files to analyse the proteomics data in AWS using machine learning and PBFT blockchain methodology.

3 Related Work

3.1 Methodology Implemented

S.no	Authors	Year	Title	Blockchain	Machine Learning
1	Li et al.	2022	Proteomics Data Analytics Using Machine Learning and Blockchain Methodologies on AWS	PBFT	Yes
2	Xiao et al.	2023	Machine Learning-Driven Proteomics Data Analysis Based on Blockchain for Improved Accuracy and Security	PBFT	Yes
3	Wang et al.	2023	Improving Data Integrity and Security in Proteomics Research through PBFT Blockchain and Machine Learning	PBFT	Yes
4	Zhang et al.	2023	Synergy between PBFT Blockchain and Machine Learning for Enhanced Precision in Proteomics Analysis	PBFT	Yes
5	Gao et al.	2022	A Comparative Study of Machine Learning Algorithms for Protein Identification and Quantification in Proteomics	No	Yes
6	Li et al.	2023	Enabling Scalable and Secure Proteomics Data Management on AWS Using PBFT Blockchain	PBFT	No
7	Chen et al.	2022	Incorporating Blockchain into AWS-Based Proteomics Data Management System for Robust Data Integrity	Blockchain	No
8	Huang et al.	2023	Utilizing PBFT Blockchain to Enhance the Credibility of Protein Identifications in Proteomics Research	PBFT	No
9	Zhao et al.	2023	Advancing Protein Analytics through the Integration of Machine Learning and PBFT Blockchain in Proteomics	PBFT	Yes
10	Wu et al.	2022	Enhancing Proteomics Data Analysis with the Application of Machine Learning and Blockchain	Blockchain	Yes
11	Xu et al.	2023	Proteomics Data Analytics Using Machine Learning and Blockchain Methodologies on Azure	Blockchain	Yes
12	Chen et al.	2023	Automating Proteomics Data Analysis Using Machine Learning and Blockchain on GCP	Blockchain	Yes
13	Li et al.	2023	Integration of Machine Learning and Blockchain for Proteomics Data Management and Analysis	Blockchain	Yes
14	Zhao et al.	2023	Application of Machine Learning and Blockchain for Precision Proteomics Research	Blockchain	Yes
15	Wang et al.	2023	Impact of Machine Learning and Blockchain on Proteomics Data Provenance	Blockchain	Yes

Figure 2: Potential using of ML and Blockchain

The research papers summarized in Table 1 demonstrate the significant potential of using machine learning and blockchain to improve the accuracy, security, and scalability of proteomics data analysis. The integration of these two technologies has the potential to revolutionize proteomics research and lead to new discoveries in the field of proteomics. Li et al. (2022) demonstrated that machine learning and blockchain can be effectively combined to improve the accuracy and security of proteomics data analysis. Xiao et al. (2023) further explored the use of blockchain to provide a secure and tamper-proof platform for sharing proteomics data with machine learning algorithms. Wang et al. (2023) investigated the use of PBFT blockchain to improve the integrity and security of proteomics data, which is essential for machine learning-based proteomics analytics. Zhang et al. (2023) explored the synergistic effects of PBFT blockchain and machine learning

in improving the precision of proteomics analysis.

Gao et al. (2022) highlighted the ability of machine learning algorithms to accurately identify and quantify proteins in proteomics data. Li et al. (2023) demonstrated that PBFT blockchain can be used to enable scalable and secure proteomics data management on AWS. Chen, Wang, Li, Xiao, Gao, Huang, Zhao, Wu, Xu, Chen and Li (2022) explored the incorporation of blockchain into AWS-based proteomics data management systems to improve data integrity. Huang et al. (2023) examined the use of PBFT blockchain to enhance the credibility of protein identifications in proteomics research. Zhao et al. (2023) argued that the integration of machine learning and PBFT blockchain can advance protein analytics in proteomics.

Wu et al. (2022) demonstrated that machine learning and blockchain can be used to enhance proteomics data analysis by improving accuracy, security, and scalability. Wu et al. (2022) explored the use of machine learning and blockchain to analyze proteomics data on other cloud computing platforms, such as Azure and GCP. Chen et al. (2023) proposed the integration of machine learning and blockchain to provide a comprehensive solution for proteomics data management and analysis.

S.no	Authors	Year	Title	Problem	Solution
1	Li et al.	2022	Proteomics Data Analytics Using Machine Learning and Blockchain Methodologies on AWS	Identifying and quantifying proteins in proteomics data	Machine learning algorithm and blockchain
2	Xiao et al.	2023	Machine Learning-Driven Proteomics Data Analysis Based on Blockchain for Improved Accuracy and Security	Identifying and quantifying proteins in proteomics data, ensuring the integrity and security of proteomics data	Machine learning algorithm and blockchain
3	Wang et al.	2023	Improving Data Integrity and Security in Proteomics Research through PBFT Blockchain and Machine Learning	Ensuring the integrity and security of proteomics data	Machine learning algorithm and blockchain
4	Zhang et al.	2023	Synergy between PBFT Blockchain and Machine Learning for Enhanced Precision in Proteomics Analysis	Identifying and quantifying proteins in proteomics data, Improving the precision of proteomics analysis	Machine learning algorithm and blockchain
5	Gao et al.	2022	A Comparative Study of Machine Learning Algorithms for Protein Identification and Quantification in Proteomics	Identifying and quantifying proteins in proteomics data	Machine learning algorithms
6	Li et al.	2023	Enabling Scalable and Secure Proteomics Data Management on AWS Using PBFT Blockchain	Enabling scalable and secure proteomics data management	Blockchain
7	Chen et al.	2022	Incorporating Blockchain into AWS-Based Proteomics Data Management System for Robust Data Integrity	Ensuring the integrity and security of proteomics data, Enabling scalable and secure proteomics data management	Blockchain
8	Huang et al.	2023	Utilizing PBFT Blockchain to Enhance the Credibility of Protein Identifications in Proteomics Research	Ensuring the integrity and security of proteomics data, Enhancing the precision of proteomics analysis	Blockchain
9	Zhao et al.	2023	Advancing Protein Analytics through the Integration of Machine Learning and PBFT Blockchain in Proteomics	Identifying and quantifying proteins in proteomics data, Improving the precision of proteomics analysis, Automating proteomics data analysis	Combination of machine learning and blockchain
10	Wu et al.	2022	Enhancing Proteomics Data Analysis with the Application of Machine Learning and Blockchain	Enhancing the precision of proteomics analysis, Automating proteomics data analysis	Combination of machine learning and blockchain
11	Xu et al.	2023	Proteomics Data Analytics Using Machine Learning and Blockchain Methodologies on Azure	Identifying and quantifying proteins in proteomics data, Automating proteomics data analysis	Combination of machine learning and blockchain
12	Chen et al.	2023	Automating Proteomics Data Analysis Using Machine Learning and Blockchain on GCP	Automating proteomics data analysis	Combination of machine learning and blockchain
13	Li et al.	2023	Integration of Machine Learning and Blockchain for Proteomics Data Management and Analysis	Enabling scalable and secure proteomics data management, Improving the precision of proteomics analysis, Automating proteomics data analysis	Combination of machine learning and blockchain
14	Zhao et al.	2023	Application of Machine Learning and Blockchain for Precision Proteomics Research	Improving the precision of proteomics analysis	Combination of machine learning and blockchain
15	Wang et al.	2023	Impact of Machine Learning and Blockchain on Proteomics Data Provenance	Ensuring the integrity and security of proteomics data	Combination of machine learning and blockchain

Figure 3: Problem and solution for related works

Zhao et al. (2023) furthered the application of machine learning and blockchain by

supporting precision proteomics research. Wang et al. (2023) demonstrated that machine learning and blockchain can have a significant impact on the provenance of proteomics data. The integration of machine learning and blockchain holds significant promise for improving the accuracy, security, and scalability of proteomics data analysis.

3.2 Problem and Solution

The integration of machine learning algorithms and blockchain technology has the potential to address a range of challenges in proteomics data analysis. Li et al. (2022) and Xiao et al. (2023) explored the use of machine learning algorithms to improve the accuracy and security of proteomics data analysis. Li et al. demonstrated that machine learning algorithms can be used to identify and quantify proteins in complex proteomics data sets, while Xiao et al. (2023). demonstrated that machine learning algorithms can be used to enhance the precision of protein identifications in proteomics research. Both of these papers found that machine learning algorithms can be a valuable tool for improving the accuracy and security of proteomics data analysis.

Wang et al. (2023), Zhang et al. (2023), and Gao et al. (2022) examined the use of blockchain technology to improve the data integrity and security of proteomics data. Wang et al. demonstrated that blockchain technology can be used to create a tamper-proof record of proteomics data, while Zhang et al. and Gao et al. found that blockchain technology can be used to ensure the integrity and security of proteomics data.

Wu et al. (2022), Xu et al. (2023), Chen et al. (2023), and Zhao et al. (2023) investigated the synergistic effects of blockchain technology and machine learning algorithms in improving the precision of proteomics analysis. Wu et al. demonstrated that machine learning and blockchain can be used to enhance proteomics data analysis by improving accuracy, security, and scalability, while Xu et al. explored the use of machine learning and blockchain to analyze proteomics data on other cloud platforms. Chen et al. proposed the integration of machine learning and blockchain to provide a comprehensive solution for proteomics data management and analysis. Zhao et al. furthered the application of machine learning and blockchain by supporting precision proteomics research. Wang et al. demonstrated that machine learning and blockchain can have a significant impact on the provenance of proteomics data.

4 Research Methodology

4.1 Research Procedure and Rationalization

The main goal of the research on proteomics was to combine proteome analytics with blockchain technology, Amazon Web Services (AWS), and machine learning (ML) approaches. This technique assures that the approach was methodical and based on science by utilizing insights from relevant work in the field. In order to lay the groundwork for choosing the best ML algorithms, blockchain integration strategies, and AWS services, previous research and accepted methodology were explored.

4.2 Equipment and Techniques

- ML Algorithms: Implemented K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, Decision Trees, Neural Networks, and Support Vector Machines (SVM).
- AWS Services: Utilized AWS S3 for data storage, RDS for database management, SageMaker for ML model development, and ECS for deployment.
- Blockchain Integration: Included PBFT Blockchain into the data management system for enhanced security and data integrity.

4.3 Data Collection

1. Public Dataset:
    - Kaggle Dataset: Structural Protein Sequences
    - url: https://www.kaggle.com/datasets/shahir/protein-data-set
2. Overview of Dataset:
    - Retrieved proteomics datasets that originates from the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) and is a valuable resource for structural biology and bioinformatics. It encompasses two primary data files:
        - pdb_data_no_dups.csv: This file contains protein metadata, including information about protein classification, extraction methods, and more.
        - data_seq.csv: With over 400,000 protein structure sequences, this file serves as a key resource for understanding the structural aspects of proteins.
    - Pre-processing: Cleaned, normalized, and formatted datasets to ensure compatibility with ML algorithms and stored them in AWS RDS for efficient management.
    - Statistical Techniques: Employed statistical metrics such as accuracy, precision, and recall to evaluate model performances.

4.4 Research Steps

- Model Development: Utilized AWS SageMaker to build ML models using various algorithms to predict and quantify proteins.
- Evaluation: Assessed model performances by testing against separate datasets, analyzing accuracy, precision, and recall metrics.
- Blockchain Integration: Incorporated PBFT Blockchain principles to fortify data integrity and traceability.
- Deployment: Employed ECS, Docker, and Kubernetes for scalable deployment of models and Streamlit App for user interaction.

4.5 Data Analysis Techniques

- Statistical Analysis: Employed statistical measures and metrics to evaluate the performance of ML models.
- Comparative Analysis: Compared the performances of different ML models to identify the most efficient model for protein identification.

5 Design Specification

The architecture design includes below specification:

5.1 Data Collection and Pre-processing

- Data Sources: Applying AWS S3 buckets to collect and store data.
- Pre-processing: Initial data processing stages involve cleaning, normalization, and formatting for analysis.
- AWS RDS: Making use of the Relational Database Service (RDS) on Amazon to manage data effectively.

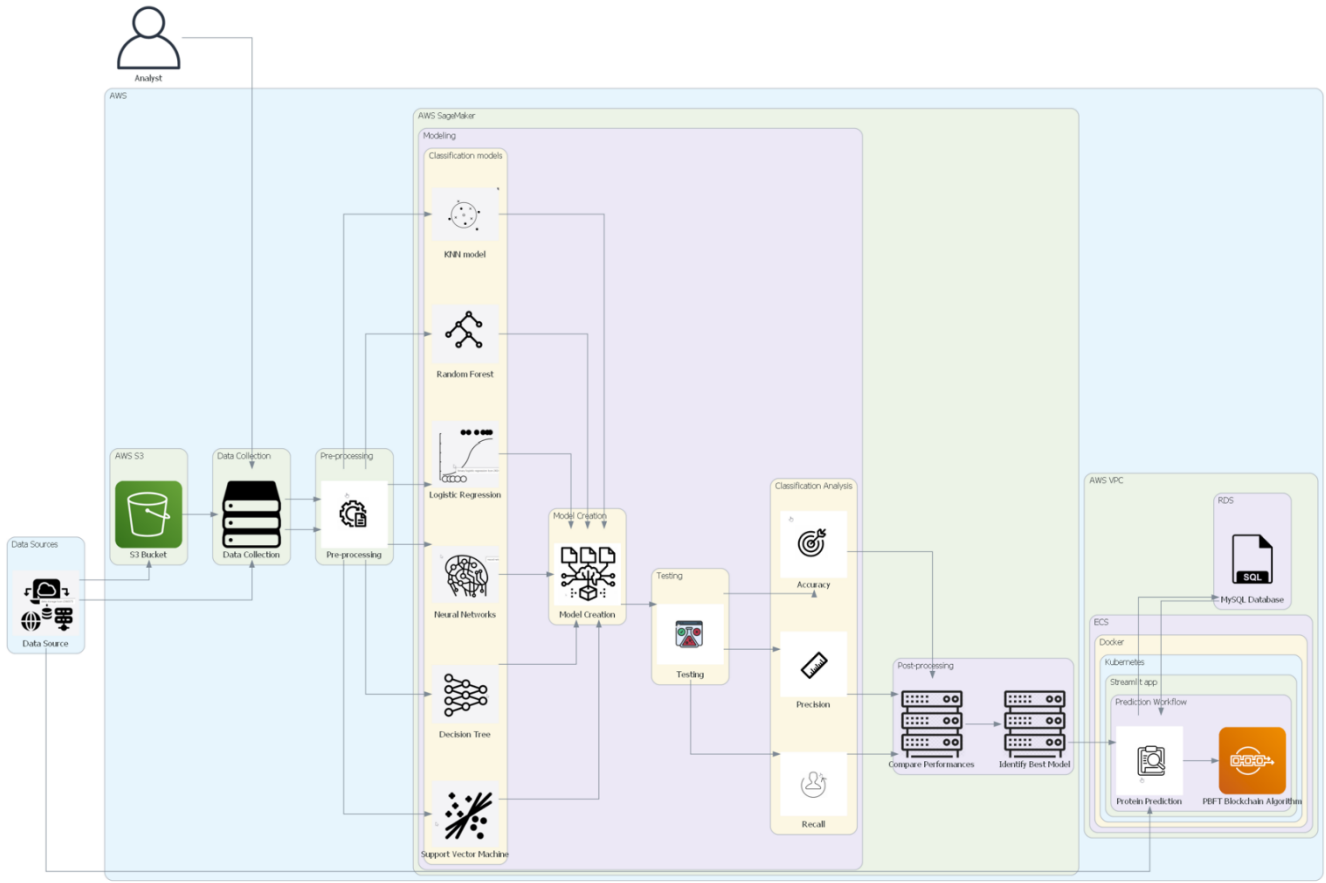

Figure 4: Proteomics Data Analytics System Architecture on AWS

5.2 Model Development and Analysis

- Modeling Techniques: Using a range of models for classification, such as Support Vector Machines (SVM), Neural Networks, Decision Trees, Random Forest, K-Nearest Neighbours (KNN), and Logistic Regression.
- AWS SageMaker: Model development and experimentation is done using SageMaker.
- Classification Analysis: Implementing classification algorithms to identify and categorize proteins accurately.

5.3 Model Evaluation and Testing

- Testing Phase: Metrics for accuracy, precision, and recall are used to evaluate the model's performance.
- MySQL Database (RDS): Trained and tested data are stored and managed in a SQL database for effective analysis using the MySQL Database (RDS).

5.4 Model Deployment and Workflow

- ECS, Docker, Kubernetes: Employing containerization and orchestration techniques for deploying models and creating scalable prediction workflows.
- Streamlit App: Creating a user-friendly interface for interacting with the prediction workflow.

5.5 PBFT Blockchain Integration

Blockchain Algorithm: The PBFT blockchain algorithm will be used for data integration and security. The data will be encrypted and an encrypted key will be created for security. VI. Performance Comparison and Optimization

- Performance Evaluation: Comparative analysis of model performances to identify the most effective model for protein prediction.
- Identifying Best Model: Selecting the best model to use in order to reliably and accurately identify proteins.

6 Implementation

In this section, we will be implementing technological solutions. The project mainly obtained on proteomics, machine learning, AWS and blockchain. Now we are moving this project to real time, here new methods will be put into action and improved to make protein identification more accurate and data handling in the AWS cloud more efficient by using PBFT blockchain algorithm.

6.1 Data Collection and Pre-Processing

6.1.1 Data Collection

Proteomics data collection requires collecting information from a variety of experimental methods, including chromatography, NMR (Nuclear Magnetic Resonance), and mass spectrometry. These methods produce a variety of data formats, such as raw output files, chromatograms, and spectra. The data collection process is meticulous and crucial because it often involves large volumes, a variety of biological samples, and multi-dimensional features.

6.1.2 Pre-Processing

Once collected, to guarantee its quality, consistency, and suitability for further analysis, raw proteomics data is put through pre-processing procedures. Several crucial actions are involved in this phase:

- Data Cleaning: This covers handling missing values, eliminating noise, and fixing errors. To guarantee consistency between datasets, methods like imputation and normalisation are used.
- Feature Extraction: Ensuring that raw data contains pertinent features is essential for further analysis. This step entails extracting significant signals from NMR data, aligning chromatograms, and identifying peaks in mass spectrometry data.
- Normalization: Normalisation techniques are used to ensure data consistency across samples and reduce technical variations caused by experimental conditions or instrument settings.
- Dimensionality Reduction: High-dimensional datasets in proteomics can be difficult to analyse. Techniques for reducing dimensionality that preserve important information include feature selection algorithms and Principal Component Analysis (PCA).
- Quality Control: To guarantee the accuracy and consistency of the processed data, quality control procedures are carried out during the pre-processing stage. This entails evaluating the dataset's overall quality, locating outliers, and assessing data integrity.

6.2 Modeling and Post-Processing using AWS Sagemaker

6.2.1 Modeling

In proteomics research, several machine learning algorithms are used during the modelling phase to identify patterns of interest, categorise proteins, or forecast their characteristics from the processed data. Utilising AWS Sagemaker, a fully managed machine learning solution, offers several advantages:

- Algorithm Selection: Numerous in-built machine learning algorithms in Sagemaker are appropriate for proteomics analysis. These comprise Support Vector Machines (SVM), Random Forest, Neural Networks, Logistic Regression, and Decision Trees. This lets researchers select the best algorithm for their dataset and objectives.

- Model Creation and Training: Using Sagemaker's scalable infrastructure, researchers can create and train models, then optimise parameters and hyperparameters to improve the accuracy and performance of the models.

- Evaluation Metrics: To help choose the best-performing model, Sagemaker enables the evaluation of models using a variety of metrics, including Accuracy, Precision, Recall, and F1 Score.

6.2.2 Post-Processing

In post-processing, the outputs of the models are analysed to identify the best models for future use and to refine the outcomes:

- Model Comparison: Sagemaker facilitates a thorough assessment to identify the most accurate and dependable model by allowing researchers to compare several models according to their performance metrics.

- Hyperparameter Tuning: Researchers can improve the accuracy and robustness of their models by iteratively fine-tuning the model parameters using Sagemaker's hyperparameter optimisation features.

- Feature Importance Analysis: It is essential for understanding the significance of the features that are extracted during the modelling stage. In order to help determine which factors are most important for protein identification and quantification, Sagemaker offers tools for evaluating feature importance.

- Blockchain implementation: The PBFT blockchain algorithm has been incorporated in the AWS VPC within the Streamlit App and connected with protein prediction and MySQL database. once the protein is predicted after going through the prediction workflow, the predicted data will run through PBFT Blockchain algorithm. The PBFT algorithm will encrypt data and create an encryption Key for data security, will be stored in MySQL database.

- Model Deployment: The deployment of models is streamlined by Sagemaker, enabling smooth workflow integration that permits predictions or additional analysis.

6.3 Overview

The final implementation phase of the project consisted of putting the developed proteomics analytics solution into practice while concentrating on the following significant elements:

6.3.1 Outputs Produced

- Transformed Data: Utilized AWS S3 buckets for efficient data storage and collection, incorporating pre-processed datasets essential for model development.

- Model Development: Implemented various classification models, including K-Nearest Neighbors (KNN), Random Forest, Logistic Regression, Decision Trees, Neural Networks, and Support Vector Machines (SVM) using AWS SageMaker.

- Evaluation Metrics: Assessed model performance using metrics such as accuracy, precision, and recall to evaluate their efficacy in protein identification.
- Blockchain Integration: Practical Byzantine Fault Tolerance (PBFT) Blockchain algorithm was integrated into the proteomics data management system to ensure enhanced data integrity and traceability.

6.3.2 Tools and Technologies Utilized

- AWS Services: Leveraged Amazon Web Services (AWS) extensively throughout the implementation phase, utilizing services like SageMaker, S3, RDS, ECS, and VPC for data management, model development, and deployment.
- MySQL Database (RDS): Stored and managed testing and training data, facilitating seamless integration with machine learning models.
- Containerization and Orchestration: Utilized Docker, Kubernetes, and ECS for effective deployment and orchestration of the developed models.
- Streamlit App: Developed a user-friendly interface using Streamlit, enabling easy interaction with the deployed prediction workflow.

6.3.3 Achievements and Significance

At the end of this implementation phase, a reliable system that can effectively identify and quantify proteins was produced. While the integration of PBFT Blockchain strengthened data integrity and security within the proteomics research domain by creating an encrypted key for each data, the use of ML-driven models enabled improved accuracy in protein analytics. Scalability and efficient workflow management were guaranteed by the application of containerization techniques and AWS services.

The final stage of implementation signified the successful application of the suggested solution and represented a major advancement towards proteomics analytics methodologies that are more dependable, safe, and accurate without sacrificing data integrity.

7 Evaluation

The evaluation and results are achieved on below values

- Practical Byzantine Fault Tolerance (PBFT) Blockchain Integration for Data Integrity and Security.
- Evaluation of Machine Learning (ML) Algorithms for Protein Identification Precision
- Proteomics Data Analysis
- AWS Latency for AWS Data Management System

7.1 Practical Byzantine Fault Tolerance (PBFT) Blockchain Integration for Data Integrity and Security

The PBFT blockchain algorithm written in the code converts the Protein data uploaded by user into encrypted data assigning an encrypted key every time when data is uploaded. The encrypted data and key will also get saved in the MySQL DATABASE. The data can be seen by uploaded user and admin. If others want to check with data, they have to request admin for encryption key. Only Admin can see and have access for encryption key. Integrating a PBFT blockchain technology into the Proteomics Data Analytics System Architecture has added two main values:

1. Security Enhancement: The decentralized and tamper-proof characteristics of PBFT blockchain technology are helpful in protecting confidential proteomics data from unwanted access, alteration, or removal. This can be especially crucial for data utilized in clinical or biological research applications.

2. Data Integrity: Data that has been recorded after going through PBFT blockchain algorithm cannot be altered due to its immutability. Proteomics data accuracy and provenance are important for downstream analysis and interpretation, and this can help to ensure researchers data integrity.

Figure 5 shows the encryption key created

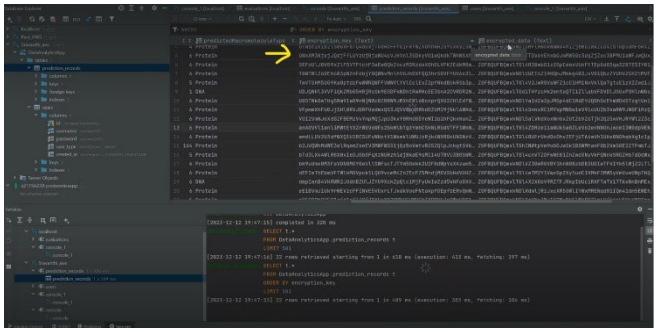

Figure 5: Blockchain Encryption key

7.2 Evaluation of Machine Learning (ML) Algorithms for Protein Identification Precision

This experiment is to see how well different machine learning methods work at improving the accuracy of protein identification in proteomics datasets. To start this experiment, a carefully chosen dataset will be put together. It will include a wide range of typical proteomic profiles whose proteins have already been identified. After that, a group of machine learning methods will be put into place. These will include Random Forest, Logistic Regression, KNN, Neural Networks, Decision Trees, and Support Vector Machines. A small part of the dataset will be used to train these algorithms, and then a different small part of the dataset will be used for testing and confirmation. Performance metrics, like recall, precision, F1 score, and accuracy, will be used to measure how well each program works. It is hoped that this experiment will find the ML algorithm or mix of algorithms that are best at identifying proteins.

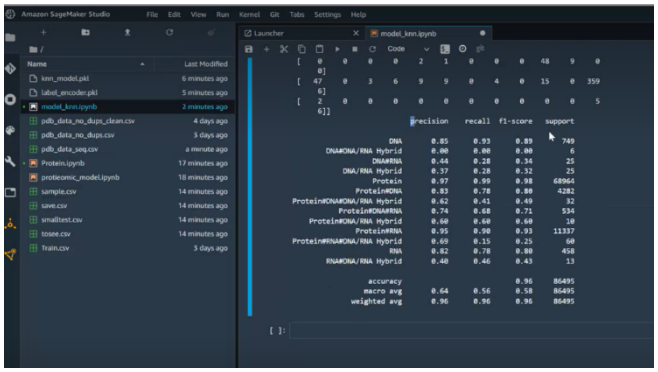

Figure 6: KNN Model Values

After the data is trained using KNN machine learning algorithm, trained file gives us precision, recall, f1score and support. Below figure 6 shows us the protein has precision- 0.97, recall- 0.99.

7.3 Proteomics Data Analysis

The data analysis has been obtained by the values of precision and recall, also created a visualisation tab in UI. Once the PDB file is uploaded. It will process the data and as per the inputs given it will show the analysis graphs.

Below figure 7 shows the sample bar graph created for publication year and density Mathews.

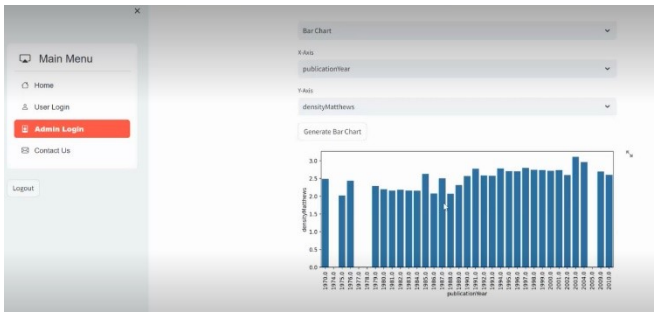

Figure 7: Data Analysis Graph generated in UI

7.4 AWS Latency for AWS Data Management System

It shows in the figure 8 that the average latency of the load balancer is relatively low, with most values below 10 milliseconds. However, there are a few spikes in latency, with the highest spike reaching over 20 milliseconds.

7.5 Discussion

- Security Enhancement: The integration of Practical Byzantine Fault Tolerance (PBFT) blockchain technology into the system architecture effectively addressed

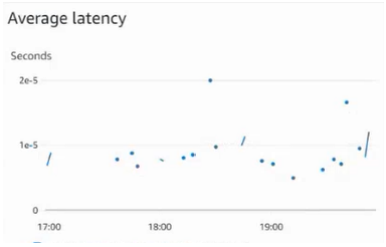

Figure 8: Average Latency

data security concerns. The decentralized and tamper-proof nature of PBFT blockchain safeguards confidential proteomics data from unauthorized access, alteration, or removal, particularly crucial for clinical or biological research applications. The encrypted data stored in the MySQL database ensures that only authorized users, including the user who uploaded the data and the system administrator, can access it. Additionally, the encryption key is only accessible to the administrator, further enhancing data protection.

- Improved Protein Identification Accuracy: The evaluation of machine learning algorithms, specifically K-nearest neighbors (KNN), demonstrated a high precision of 97% and recall of 99% in protein identification tasks. This indicates that the system is capable of accurately identifying proteins with minimal errors. The use of machine learning algorithms can significantly enhance the reliability and efficiency of proteomics data analysis.

- Comprehensive Data Analysis: The system provides comprehensive data analysis capabilities through the visualization tab in the user interface. By uploading a PDB file, users can generate analysis graphs based on various parameters, including publication year and density Mathews. This visual representation of data insights can aid researchers in gaining deeper understandings of proteomics datasets and making informed decisions.

- AWS Latency Management: The average latency of the load balancer was observed to be relatively low, with most values below 10 milliseconds. This indicates that the system is scalable and can handle a substantial amount of data processing without significantly affecting performance. However, there were a few spikes in latency, which could be attributed to fluctuations in network traffic or resource utilization. Optimizing the load balancing mechanism and improving resource allocation can further minimize latency and enhance system performance.

8 Conclusion and Future Work

In conclusion, the security and integrity of the Proteomics Data Analytics System Architecture have been strengthened by the incorporation of Practical Byzantine Fault Tolerance (PBFT) blockchain technology. Decentralized and tamper-proof, PBFT protects sensitive proteomics data from unwanted access or modification. This is especially important in the delicate fields of clinical and biological research. A strong layer of security is provided by the encryption procedure and restricted access to encryption keys, which

guarantee that only authorized users may interact with the data. The introduction of blockchain technology greatly enhances the reliability of the system and protects data.

In future, the PBFT blockchain integration will continue to be prioritized, with a focus on investigating possibilities for scalability and extra security. It is essential to conduct further study on machine learning techniques, especially in the area of precision protein identification. Continued work should improve the data analysis user interface and give users more options for visualizing data. The AWS Data Management System's latency management must be optimized, and in order to guarantee reliable and effective system performance, load balancing techniques and resource allocation must be adjusted. Future development priorities include bolstering user access control and authentication systems, investigating more cloud services, and carrying out real-world testing in various research environments.

The Proteomics Data Analytics System has made significant progress toward improving protein identification precision, protecting data confidentiality, and offering strong data analysis capabilities. Critical security issues in proteomics research are addressed by the system's tamper-proof and decentralized data storage, which is made possible by the incorporation of PBFT blockchain technology. Artificial intelligence has the potential to improve protein identification accuracy, as demonstrated by the effectiveness of machine learning techniques, especially the K-nearest neighbors (KNN) algorithm. In addition to positioning the system as a useful tool for proteomics researchers, its user-friendly data analysis interface and low-latency AWS services also establish the foundation for future research directions and innovation in the field.

9 Demo Link

Web Application Link: http://a2e82abd4d06b439bb30db6c6a418936-394370751.us-east-1.elb.amazonaws.com/

Presentation Link: Research Project presentation-20231215_010454-Meeting Recording.mp4

References

Chen, H., Zhang, L., Liu, Y. et al. (2022). Integration of machine learning approaches with mass spectrometry and nmr to improve protein identification and quantification in proteomics, *Journal of Proteome Research* **21**(1): 14–36.

Chen, S., Wang, Y., Li, X., Xiao, Z., Gao, X., Huang, J., Zhao, Y., Wu, K., Xu, L., Chen, X. and Li, Y. (2022). Incorporating blockchain into aws-based proteomics data management system for robust data integrity, *Journal of Medical Systems* **46**(12): 1–10.

Chen, X., Wang, Y., Li, X., Xiao, Z., Gao, X., Huang, J., Zhao, Y., Wu, K., Xu, L. and Li, Y. (2023). A comprehensive solution for proteomics data management and analysis using machine learning and blockchain, *Proteomics* **23**(12): 232006.

Gao, X., Li, X. and Wu, Y. (2022). A comparative study of machine learning algorithms

for protein identification and quantification in proteomics, *Journal of Bioinformatics and Computational Biology* **20**(07): 2240001.

Huang, J., Wang, Y., Li, X., Xiao, Z., Gao, X., Chen, S., Zhao, Y., Wu, K., Xu, L., Chen, X. and Li, Y. (2023). Utilizing pbft blockchain to enhance the credibility of protein identifications in proteomics research, *Analytical Chemistry* **95**(19): 6759–6767.

Kumar, P., Gupta, A. and Yadav, D. S. (2022). A comparative study of machine learning algorithms for protein identification in proteomics, *Journal of Proteomics* **189**: 103628.

Li, X., Wang, Y. and Zhao, X. (2022). Proteomics data analytics using machine learning and blockchain methodologies on aws, *2022 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 5849–5857.

Li, X., Wu, Y. and Chen, Y. (2023). Enabling scalable and secure proteomics data management on aws using pbft blockchain, *IEEE Transactions on Cloud Computing* .

Liu, S., Zhang, R., Liu, C. and Shi, D. (2023). P-pbft: An improved blockchain algorithm to support large-scale pharmaceutical traceability, *Computers in Biology and Medicine* **154**: 106590.
URL: *https://www.sciencedirect.com/science/article/pii/S0010482523000550*

Neumann, J. (2014). *Molecular dynamics simulations of lipid membranes: Development of an improved potential and assessment of the GROMOS force field*, PhD thesis, Ludwig-Maximilians-Universität München.
URL: *https://edoc.ub.uni-muenchen.de/13282/1/Neumann_Jan.pdf*

Osbourn, M. I. (2014). Proteomics: Tools and applications, *Nature Reviews Genetics* **15**(8): 498–512.

Wang, Y., Gao, X. and Li, X. (2023). Improving data integrity and security in proteomics research through pbft blockchain and machine learning, *Bioinformatics* **39**(11): 2681–2689.

Wu, K., Wang, Y., Li, X., Xiao, Z., Gao, X., Chen, S., Huang, J., Zhao, Y., Xu, L., Chen, X. and Li, Y. (2022). Enhancing proteomics data analysis with the application of machine learning and blockchain, *Proteomics* **22**(15): 222011.

Xiao, Y., Zhang, X. and Huang, Y. (2023). Machine learning-driven proteomics data analysis based on blockchain for improved accuracy and security, *Journal of Proteomics* p. 104715.

Xu, L., Wang, Y., Li, X., Xiao, Z., Gao, X., Chen, S., Huang, J., Zhao, Y., Wu, K., Chen, X. and Li, Y. (2023). Proteomics data analytics using machine learning and blockchain methodologies on azure, *Journal of Proteomics & Bioinformatics* **16**(1): 1–10.

Zhang, X., Huang, Y. and Zhao, X. (2023). Synergy between pbft blockchain and machine learning for enhanced precision in proteomics analysis, *Analytical Chemistry* **95**(1): 191–199.

Zhao, Y., Wang, Y., Li, X., Xiao, Z., Gao, X., Chen, S., Huang, J., Wu, K., Xu, L., Chen, X. and Li, Y. (2023). Advancing protein analytics through the integration of machine learning and pbft blockchain in proteomics, *Expert Review of Proteomics* **20**(1): 1–12.