Implementation of Machine Learning for Breast Cancer Classification Based on Genomic Data: Backend Solution with Supabase and Streamlit

1st Tia Hasna Humayra
Faculty of Electrical Engineering
Telkom University
Bandung, Indonesia
tiasna@student.telkomuniversity.ac.id

2nd Suryo Adhi Wibowo
Faculty of Electrical Engineering
Telkom University
Bandung, Indonesia
suryoadhiwibowo@telkomuniversity.ac.id

3rd Koredianto Usman
Faculty of Electrical Engineering
Telkom University
Bandung, Indonesia
korediantousman@telkomuniversity.ac.id

Abstrak — Kanker payudara tetap menjadi salah satu penyebab utama kematian terkait kanker di seluruh dunia, menyoroti kebutuhan akan alat diagnostik yang akurat dan efisien. Studi ini berfokus pada penerapan model pembelajaran mesin, khususnya Jaringan Saraf Tiruan (ANN), untuk mengklasifikasikan jenis kanker payudara berdasarkan data genomik. Menggunakan dataset METABRIC RNA Mutation, sistem ini menggabungkan backend berbasis cloud dengan Supabase dan frontend intuitif yang dibangun dengan Streamlit. Untuk memastikan kompatibilitas data dengan model, langkah-langkah pra-pemrosesan seperti standardisasi, pengkodean label, dan pengkodean one-hot diterapkan. TensorFlow digunakan untuk memuat model yang disimpan dalam format .h5, dengan dua pendekatan yang diuji: model dengan 30 fitur yang mencapai akurasi 99% dan waktu prediksi rata-rata 80 milidetik, serta model dengan 6 fitur yang mencapai akurasi 100% dengan waktu prediksi yang lebih cepat yaitu 42,25 milidetik. Hasil prediksi disimpan dengan aman di Supabase, lengkap dengan cap waktu untuk pelacakan dan diekspor sebagai laporan PDF untuk dokumentasi yang mudah. Keamanan data diprioritaskan melalui penggunaan kunci API, token JWT, dan manajemen rahasia Streamlit untuk melindungi informasi sensitif. Integrasi Supabase untuk pemrosesan backend, Streamlit untuk visualisasi waktu nyata, dan GitHub untuk otomatisasi CI/CD menghasilkan sistem yang skalabel, andal, dan efisien. Studi ini menyajikan solusi yang kuat untuk klasifikasi kanker payudara, menyediakan prediksi waktu nyata, penanganan data yang aman, dan antarmuka yang ramah pengguna yang cocok untuk aplikasi klinis dan penelitian.

***Kata kunci*— klasifikasi kanker payudara, *artificial neural network*, data genomik, Supabase, Streamlit, prediksi *real time*, keamanan data.**

***Abstract*— Breast cancer remains one of the leading causes of cancer-related deaths worldwide, highlighting the need for accurate and efficient diagnostic tools. This study focuses on implementing machine learning models, particularly Artificial Neural Networks (ANN), to classify breast cancer types based on genomic data. Using the METABRIC RNA Mutation dataset, the system combines a cloud-based backend with Supabase and an intuitive frontend built with Streamlit. To ensure data compatibility with the models, preprocessing steps such as standardization, label encoding, and one-hot encoding are applied. TensorFlow is used to load models saved in .h5 format, with two approaches tested: a 30-feature model achieving 99% accuracy and an average prediction time of 80 milliseconds, and a 6-feature model achieving 100% accuracy with a faster prediction time of 42.25 milliseconds. Prediction results are stored securely in Supabase, complete with timestamps for tracking and exported as PDF reports for easy documentation. Data security is prioritized through the use of API keys, JWT tokens, and Streamlit secret management to safeguard sensitive information. The integration of Supabase for backend processing, Streamlit for real-time visualization, and GitHub for CI/CD automation results in a scalable, reliable, and efficient system. This study presents a robust solution for breast cancer classification, providing real-time predictions, secure data handling, and a user-friendly interface suitable for clinical and research applications.**

***Keywords*— breast cancer classification, artificial neural network, genomic data, Supabase, Streamlit, real-time prediction, data security.**

I. INTRODUCTION

In 2019, breast cancer had a significant number in the United States; as many as 268,000 cases of invasive breast cancer were found in women, and as many as 62,930 women in the United States were diagnosed with invasive breast cancer. The death rate due to breast cancer reaches 42,000 women every year [1]. Breast cancer is a type of cancer that develops in breast cells and occurs when the cells begin to grow abnormally and uncontrolled [2]. Breast cancer treatment is carried out by chemotherapy; many women experience high anxiety due to breast cancer [1]. Several factors that cause breast cancer risk are family history, lack of physical activity, and psychological stress [3]. Breast cancer is the second cause of cancer death among women after lung cancer [3]. Breast cancer detection is essential to understand the type of cancer and how to treat it using machine learning.

Early detection of breast cancer can use machine learning [3]. Machine learning is a science of computers that allows computers to learn data themselves without being explicitly programmed [4]. Classification is categorized into several approaches: supervised, unsupervised, semi-supervised, and reinforcement learning [4]. Several machine learning models can be used to classify breast cancer, such as random forest, support vector machine (SVM), and artificial neural network (ANN) [5]. Machine learning is used to classify cancer patients, and classification is based on features extracted from training data [6].

In this research, the website was used as a platform to detect the type of breast cancer, and the website application development is divided into two parts: front-end and back-end. The front end is the part of the web application that interacts directly with the user, where it is used to display information to the user and handle user interactions [7]. The back-end includes servers, applications, and databases that work behind the scenes [7].

An open-source platform called Supabase is used to create backend apps. A full Postgres database is included with every Supabase project, and tables may be made using the Supabase dashboard. A secure API that allows CRUD operations is automatically generated by Supabase. Additionally, Supabase offers a number of user authentication options, including password and email. [8]. This development focuses mainly on the backend, which uses Supabase as a cloud-based platform and Streamlit to run machine learning models in real-time.

II. RELATED WORK

Several studies often carry out breast cancer analysis using cloud computing [9] . Cloud computing is much frequented to analyze breast cancer, especially in handling large and complex genetic data such as DNA, RNA, and gene expression [9]. Several existing researchers use cloud computing to analyze breast cancer types because cloud computing has many functions, such as providing flexible and scalable resources [10]. It also has fast processing with machine learning, such as using a random forest model [10]. Then, easy accessibility and collaboration can work together using the same platform and allow easy collaboration, real-time data sharing, and compassionate medical data; the cloud provides a high level of security with encryption and access control, which can maintain the confidentiality of patient data [10].

Several studies often conduct breast cancer analysis using machine learning [11]. The use of machine learning in detecting breast cancer can improve diagnostic accuracy [11]. The use of algorithms such as Support Vector Machine (SVM), decision tree, and Artificial Neural Network (ANN) in predicting breast cancer recurrence, where the Support Vector Machine (SVM) model provides the highest accuracy of 95.7% [11]. This research explains the importance of selecting the right features and the appropriate processing model to improve diagnostic results [11].

In research on Scalable Pathogen Pipeline Platform (SP3), cloud computing is used in the medical field, including breast cancer analysis. It was explained that cloud computing makes it possible to analyze gene data on a large scale. The Pathogen Pipeline Platform (SP3) can utilize cloud computing to run complex bioinformatics workflows using the scalable platform. This platform enables efficient analysis of pathogen genes using the cloud. Cloud computing provides tools that can be accessed and used on various infrastructures, such as Google Cloud and Microsoft Azure [12].

The paper Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics explained that they use cloud computing for breast cancer analysis. Cloud computing can be integrated with artificial intelligence and radiomics technology to predict breast cancer response for treatments such as chemotherapy. This technology can process extensive medical data, such as magnetic resonance imaging (MRI) images and feature extraction. It is important to assist doctors in making quick and accurate decisions. Cloud computing can store, process, and analyze data on a large scale, which supports real-time medical data analysis without requiring additional hardware [13].

Distributed Intrusion Detection System using Blockchain and Cloud Computing Infrastructure research uses the Amazon Web Services (AWS) and Google Cloud Platform (GCP) platforms to provide large-scale data storage, analysis, and processing services. Machine learning technology can help diagnose and predict medically based diseases, such as in research, and analyzing breast cancer patients' responses [14].

According to Ayezabu Amanuel's research, Supabase is a PostgreSQL database that is useful for complex queries and allows self-hosting by giving infrastructure developers more control. Price analysis shows that Supabase offers a free tier with unlimited API calls. Supabase is one of the platforms for building Progressive Web Apps (PWAs), with a focus on performance evaluation, features, pricing, and stability [8].

In general, although several studies use cloud computing in breast cancer analysis, in-depth research in this study was carried out by analyzing the type of breast cancer suffered by patients, which was integrated using a dashboard that would visualize the results of the analysis, as well as real-time processing parameters which would automatically be updated when there is new data. This is an essential feature in the medical world in making quick and correct decisions in analyzing the type of breast cancer suffered by the patient.

III. MATERIAL AND RESEARCH METHO

A. *Model*

The model used in this implementation is the Artificial Neural Network (ANN) to classify types of breast cancer based on genomic data. This research uses the METABRIC RNA Mutation dataset [15] with format ".csv" which contains genetic, histologic, and breast cancer classification features.

The initial stage includes data preprocessing, where missing values are filled or removed, categorical features are processed using one-hot encoding or label encoding, and numerical data is standardized using a standard scaler. In the feature selection stage, two approaches are used, namely 30 features shown in Table 1, with an accuracy 99%, and 6 features shown in Table 2, with an accuracy of 100%. The ANN model is built using TensorFlow Keras. After training, the model is saved in ".h5" format. this model is implemented in a web-based application using Streamlit. The backend processes the input data first before passing it to the ANN model to generate breast cancer type classifications.

TABLE I. 30 FEATURES WITH ONE TARGET

Features	**Data type**
neoplasm_histologic_grade	float64
aurka	float64
chek1	float64
ccne1	float64
ahnak	float64
e2f2	float64
cdc25a	float64
aph1b	float64
cdh1	float64
gsk3b	float64
lama2	float64
src	float64
tgfb3	float64
slc19a1	float64
chemotherapy	category

lfng	float64
mapt	float64
cdk1	loat64
hsd17b10	float64
bcl2	float64
oncotree_code	category
tumor_other_histologic_subtype	category
ahnak2_mut	category
kmtd2_mut	category
stab2_mut	category
pde4dip_mut	category
map3k1_mut	category
muc16_mut	category
cdh1_mut	category
atr_mut	category
cancer_type_detailed	category

TABLE II. 6 FEATURES WITH ONE TARGET

Features	Data type
tumor_other_histologic_subtype	category
oncotree_code	category
ahnak2_mut	float64
aurka	float64
ccne1	float64
src	float64
cancer_type_detailed	category

B. *Implementation of Cloud Computing*

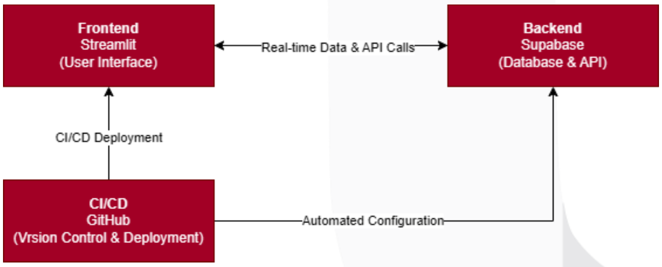


Figure 1. Cloud Computing Workflow

Figure 1 represents a cloud computing workflow integrating Streamlit (Frontend), Supabase (Backend/Database), and GitHub (CI/CD) to create an efficient, secure, and automated system. The front end, built with Streamlit, serves as the user interface where users can interact with the system. Users perform actions such as signing up, logging in, inputting data, and viewing prediction results. These actions trigger API calls that connect the front end to the back end. The backend, powered by Supabase, processes these requests by handling user authentication (using JWT tokens), validating inputs, executing ML and DL models, and storing or retrieving data in its database.

To ensure seamless integration and deployment, GitHub is utilized for version control and CI/CD pipelines. Each time a code update is pushed to the repository, the CI/CD workflow automatically tests, validates, and deploys the latest version of the application. This ensures that both the front end and back end remain up to date with minimal manual intervention. The interaction between the components highlights a streamlined process: Streamlit captures user inputs, Supabase processes and secures the data, and GitHub automates deployments. This architecture not only ensures real-time data processing but also emphasizes scalability, security, and maintainability.

C. *System Pipeline*

The flow diagram shown in figure 2 represents the main stages for the implementation of breast cancer classification. This pipeline consists of several main stages starting from user input, where users must go through the sign-up or login process integrated using bcrypt and JWT. After that, users can fill in genomic data or related information through the provided interface. In the data input, users must fill in all the available fields, which is part of the data validation process. After the data is declared valid, the next step is preprocessing, which includes data normalization. The processed data then enters the model execution stage where the machine learning model is run to classify the type of breast cancer based on the available genomic features. After that, the final result, which is the classification of the type of breast cancer, will be displayed through the Streamlit dashboard.

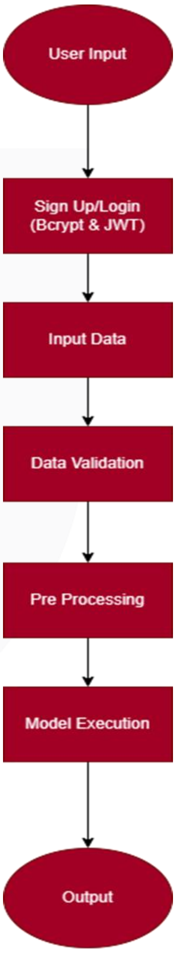


Figure 2. System Pipeline Backend

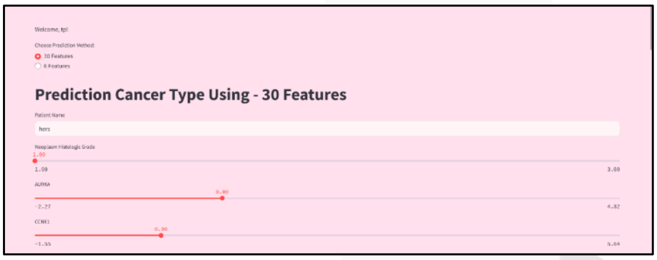
This pipeline reflects the workflow of a backend system integrated with Supabase for data management and authentication, as well as Streamlit for result visualization.

Each stage in the pipeline is designed to ensure efficiency, security, and accuracy of results, supporting the implementation of a reliable and user-accessible breast cancer classification system.

IV. RESULT AND DISCUSSION

Data processing in the system ensure that the user-entered data meets the required format for the prediction model. This process begins with users completing an input from the website which includes numerical input through sliders like “aurka”, and categorical input using dropdown menus “oncotree_code”, and text field must be filled with the patient’s name before proceeding to the detection stage that is shown in Figure 3, the website also provides 2 methods which are 30 features and 6 features.

For the machine learning model, TensorFlow is used to load models saved in the “.h5” format via the TensorFlow Keras function. Additionally, preprocessing tools such as a label encoder, standard scaler, and one hot encoder are essentials for preparing input data before go to the model. Once the model and preprocessing tools are set up, the user provided input data goes through processing. During this stage, numerical features are standardized, and categorical features are transformed into numerical values sing one hot encoder. The processed data is then combined and passed into the model using the predict function, which generates predictions based on prepared input. The prediction labels are converted back into their original format using the inverse transform method from the label encoder so that the user can understand the result. Final predictions are presented using “st.write” element through the streamlit interface.



Welcome, tpl!

Choose Prediction Method:

● 30 Features
○ 6 Features

Prediction Cancer Type Using - 30 Features

Patient Name

hers

Neoplasm Histologic Grade

1.00

1.00 3.00

AURKA

0.00

-2.27 4.82

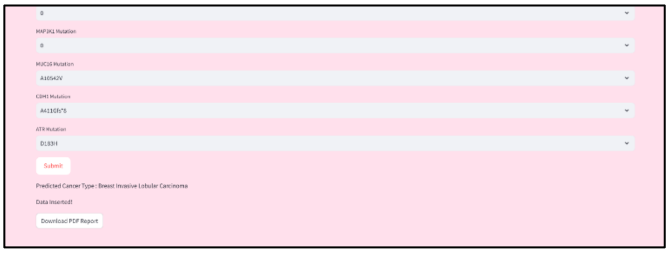
CCNE1

0.00

-1.55 5.64

Figure 3. Input Data from User

The streamlit application connects seamlessly with the supabase database using an API configured with a specified API key and URL. The connection is established using the create client function from the supabase library. Authentication is implemented through JWT tokens to manage user access permissions. After data is processed and predictions are generated, the result will be stored in the supabase tables using the insert method. Integration with the Supabase Database was successfully carried out as shown in Figure 4, a notification will appear if the data is successfully entered into the database, Figure 4 also shows the predicted result and the website also provide PDF result. The input data and prediction results are stored in the appropriate table, complete with timestamps and the correct table format as shown in Figure 5, where the database table contains various types of inputs along with the timestamps to prove that the data entering the database is done in real time and the results are immediately available in the database without delay.



0

MAP3K1 Mutation

0

MUC16 Mutation

A10542V

CDH1 Mutation

A411Gfs*8

ATR Mutation

D183H

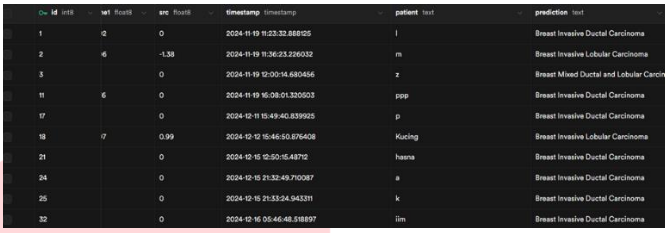
Submit

Predicted Cancer Type : Breast Invasive Lobular Carcinoma

Data Inserted!

Download PDF Report

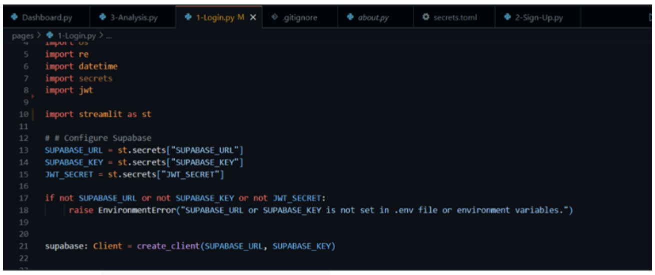
Figure 4. Result Prediction and Inserted Data to Database Notifications



id int8	ne1 float8	src float8	timestamp timestamp	patient text	prediction text
1	2	0	2024-11-19 11:23:32.888125	l	Breast Invasive Ductal Carcinoma
2	6	-1.38	2024-11-19 11:36:23.226032	m	Breast Invasive Lobular Carcinoma
3		0	2024-11-19 12:00:14.680456	z	Breast Mixed Ductal and Lobular Carcin
11	6	0	2024-11-19 16:08:01.320503	ppp	Breast Invasive Ductal Carcinoma
17		0	2024-12-11 15:49:40.839925	p	Breast Invasive Ductal Carcinoma
18	7	0.99	2024-12-12 15:46:50.876408	Kucing	Breast Invasive Lobular Carcinoma
21		0	2024-12-15 12:50:15.48712	hasna	Breast Invasive Ductal Carcinoma
24		0	2024-12-15 21:32:49.710087	a	Breast Invasive Ductal Carcinoma
25		0	2024-12-15 21:33:24.943311	k	Breast Invasive Ductal Carcinoma
32		0	2024-12-16 05:46:48.518897	iim	Breast Invasive Ductal Carcinoma

Figure 5. Data Tables from Database

To ensure data security, the website utilizes the streamlit secret management feature to store sensitive information and configure the application without exposing confidential data to the public. Security and data integrity are further maintained by sending data through an API key as illustrated in Figure 6.



```
import os
import re
import datetime
import secrets
import jwt

import streamlit as st

# # Configure Supabase
SUPABASE_URL = st.secrets["SUPABASE_URL"]
SUPABASE_KEY = st.secrets["SUPABASE_KEY"]
JWT_SECRET = st.secrets["JWT_SECRET"]

if not SUPABASE_URL or not SUPABASE_KEY or not JWT_SECRET:
    raise EnvironmentError("SUPABASE_URL or SUPABASE_KEY is not set in .env file or environment variables.")


supabase: Client = create_client(SUPABASE_URL, SUPABASE_KEY)
```

Figure 6. Data Security Using Secret

The machine learning model with 30 features was thoroughly tested to ensure the validity, performance, and efficiency in processing predictions. The result confirmed that the model was successfully loaded and compiled. The model was tested 4 times with the first trial achieving 81 ms, the second trial achieving 80 ms, the third trial achieving 74 ms, and the last trial achieving 85 ms, we can conclude that the prediction results being generated in under 1 second, achieving an average execution time of 80 milliseconds per prediction. Similarly, the model with 6 features was tested 4 times as shown in Tale 4, the average time required to load the model was 42.25 milliseconds, which is notably faster compared to the 30 features model.

TABLE III. RESULTS OF 30 FEATURES AND 6 FEATURES TRIALS

30 Features	6 Features
81 ms	38 ms
80 ms	37 ms
74 ms	54 ms
85 ms	40 ms

V. CONCLUSION

The implementation of machine learning models for breast cancer classification, integrated with streamlit and supabase. Data processing ensures that user provided inputs are validated and formatted correctly using preprocessing tools such as label encoder, standard scaler, and one hot encoder. This ensures that numerical features are standardized and categorical features are converted into numerical values for seamless integration into the machine learning pipeline.

The system successfully integrates tensorflow for loading and utilizing models that are saved in ".h5" format. Both the 30 features and 6 features models performed effectively, with prediction results generated in real time. Testing of the 30 features model achieved an average execution time of 80 milliseconds per prediction, while the 6 features model achieved an average execution time of 42.42 milliseconds. These results confirm the system's ability to handle data inputs efficiently.

Additionally, data security and integrity are maintained through the use of the streamlit secret management feature, API key, and JWT tokens, ensuring sensitive information is protected from unauthorized access. The integration with the supabase database allows for real time storage of prediction results, somplete with timestamps. The system offers a user friendly interface via streamlit, with predictions presented clearly and the option to download resuults in PDF format.

PREFERENCES

[1] A. H. A. A.-B. Arwa Okaidat, "Breast Cancer and Anxiety: A Relationship Study," in *International Conference on Information and Communication Systems (ICICS)*, 2020.

[2] A. B. S. K. B. E. A. Akhil Kumar Das, "Introduction to Breast Cancer and Awareness," in *International Conference on Advanced Computing & Communication Systems (ICACCS)*, 2021.

[3] S. D. M. H. E. H. Rahmanul Hoque, "Breast Cancer Classification using XGBoost," in *International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2020.

[4] R. G. Thomas Rincy N, "A Survey on Machine Learning Approaches and Its Techniques:," in *International Students' Conference on Electrical,Electronics and Computer Science (SCEECS)*, India, 2020.

[5] P. K. T. E. Ebru Aydındag Bayrak, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," in *Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, Turkey, 2019.

[6] S. T. Poonam Kathale, "Breast Cancer Detection and Classification," in *International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, India, 2020.

[7] F. G. K. C. F. W. Yifei Gong, "The Architecture of Micro-services and the Separation of Frond-end and Back-end Applied in a Campus Information System," in *International Conference on Advances in Electrical Engineering and Computer Applications(AEECA)*, US, 2020.

[8] A. Amanuel, "Supabase vs Firebase: Evaluation of performancand development of Progressive Web Apps," *Metropolia*, p. 58, 2022.

[9] A. A. A. A. K. A. D. P. A. Sujan Ray, "Selecting Features for Breast Cancer Analysis and Prediction," in *International Conference on Advances in Computing and Communication Engineering (ICACCE)*, USA, 2020.

[10] P. K. Tanaya Padhi, "Breast Cancer Analysis Using WEKA," in *International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Uttar Pradesh, 2019.

[11] in *International Conference on Emerging Trends in Communication, Control and Computing (ICONC3)*, 2020.

[12] D. V. P. F. J. S. M. B. S. H. Fan Yang-Turner, "Scalable Pathogen Pipeline Platform (SP^3): Enabling Unified Genomic Data Analysis with Elastic Cloud Computing," in *International Conference on Cloud Computing (CLOUD)*, UK, 2019.

[13] M. E. A. M. B. Yassine Amkrane, "Towards Breast Cancer Response Prediction using Artificial Intelligence and Radiomics," in *International Conference on Cloud Computing and Artificial Intelligence: Technologies and Applications (CloudTech)*, Belgium, 2020.

[14] A. K. S. Manish Kumar, "Distributed Intrusion Detection System using Blockchain and Cloud Computing Infrastructure," in *International Conference on Trends in Electronics and Informatics (ICOEI)*, India, 2020.

[15] R. Alharbi, "Kaggle," 2020. [Online]. Available: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric.