**MACHINE LEARNING PROJECT – CRN 30460**

**GENETIC BASED DISEASE PREDICTION AND IDENTIFICATION WITH MACHINE LEARNING**

**Team Members:**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| SAI KUMAR REDDY KETHIREDDY | 700744628 |
| SATHWIK NITTURI | 700745195 |
| SAI TEJA EDIKUDA | 700745893 |

**PROPOSAL**

Genetic-based disease prediction using machine learning has emerged as a promising approach in biomedical research. This abstract provides an overview of the importance, challenges, and advancements in this field.

Identifying disease predisposition and risk factors from genetic data plays a crucial role in personalized medicine and disease prevention. With the advent of large-scale genomic datasets, machine learning techniques have been applied to predict disease susceptibility based on genetic information. However, this task remains challenging due to the complex nature of genetic data and the need for accurate and robust predictive models.

In recent years, significant progress has been made in developing machine learning algorithms for genetic-based disease prediction. These algorithms utilize various computational techniques, such as feature selection, dimensionality reduction, and model optimization, to extract meaningful patterns and associations from genetic data. They aim to identify genetic variants or patterns that are associated with specific diseases or disease traits.

One of the key challenges in genetic-based disease prediction is the high-dimensional nature of genetic data. The human genome consists of millions of genetic variants, making it computationally demanding to analyze and interpret this vast amount of information. To overcome this challenge, feature selection methods are employed to identify the most relevant genetic markers that contribute to disease prediction. Dimensionality reduction techniques, such as principal component analysis (PCA) and t distributed stochastic neighbor embedding (t-SNE), are also utilized to reduce the dimensionality of the data while preserving important information.

Another important aspect in genetic-based disease prediction is the choice of machine learning algorithms. Various algorithms, including support vector machines (SVM), random forests, and deep learning approaches, have been applied to predict disease outcomes based on genetic data. These algorithms differ in their ability to handle high-dimensional data, capture complex interactions between genetic variants, and handle different types of diseases (e.g., binary classification, multi-class classification, or regression).

Furthermore, the availability of large-scale genomic datasets, such as the UK Biobank and the Genotype-Tissue Expression (GTEx) project, has enabled the development of more accurate and robust predictive models. These datasets provide valuable resources for training and validating machine learning models, as they contain genetic and clinical data from a diverse population.

In conclusion, genetic-based disease prediction using machine learning holds great promise in advancing our understanding of the genetic basis of diseases and facilitating personalized medicine. The integration of advanced computational techniques, large-scale genomic datasets, and diverse machine learning algorithms has significantly improved the accuracy and reliability of disease prediction models. However, further research is still needed to address the challenges associated with genetic data analysis, model interpretability, and generalization to diverse populations.

**Main Contributions and Objectives:**

**Contribution:**

* The application utilizes machine learning in Python to contribute to the identification of genetic diseases.
* The application aims to detect genetic diseases based on patient record attributes.
* It provides a helpful tool for finding and diagnosing genetic diseases using machine learning techniques.
* The application's use of machine learning algorithms enhances the accuracy and efficiency of genetic disease identification.
* By leveraging patient data and applying machine learning, the application aids in early detection and timely treatment interventions for genetic diseases.

**Objectives:**

* The project's objective is to utilize machine learning techniques to detect genetic diseases early based on available attributes.
* We will preprocess the dataset, removing noise and null values.
* We will conduct data analysis and visualization to facilitate further processing.
* We will select a suitable machine learning algorithm for accurate disease prediction.
* We will divide the dataset into training, testing, and validation sets, which will include brain images associated with genetic diseases.

**Motivation :**

1. Unveiling Hidden Insights: Genetic-based disease identification and prediction offer an opportunity to uncover valuable insights into the underlying genetic factors contributing to diseases. This can provide crucial knowledge about disease mechanisms, genetic variations, and potential therapeutic targets, advancing our understanding of human health and biology.

2. Personalized Medicine and Treatment: Genetic-based disease identification allows for personalized medicine approaches, tailoring treatments based on an individual’s unique genetic profile. This can lead to more precise and effective interventions, reducing the risk of adverse reactions and improving treatment outcomes. By customizing healthcare strategies, we can optimize patient care and enhance the overall quality of healthcare delivery.

3. Impact on Public Health and Well-being: Genetic diseases impose a significant burden on individuals, families, and healthcare systems. By focusing on genetic-based disease identification and prediction, we can make a substantial impact on public health. Early detection and intervention can help in implementing preventive measures, reducing disease prevalence, and minimizing the socioeconomic burden associated with genetic disorders, ultimately improving the well-being of affected individuals and society as a whole.

Facts: - Genetic diseases affect a substantial portion of the population, with over 4,000 identified genetic disorders. - Early detection and intervention can significantly improve patient outcomes and reduce healthcare costs associated with managing genetic diseases. - Genetic-based disease identification has the potential to transform healthcare by enabling personalized medicine approaches and targeted treatments based on an individual’s genetic profile. - Through genetic-based prediction and preventive measures, it is possible to reduce the incidence and impact of genetic diseases, improving public health and well-being.

**Significance:**

The project development will utilize Google Colab, which is a Python tool, enabling its execution on any computer system with an internet connection. This approach eliminates the requirement for specific software installations on the user's system. The Colab tool facilitates the development and execution of the application directly within a cloud server where the required Python library files are installed. The machine learning algorithm libraries are also integrated into Colab, enabling the project to utilize machine learning algorithms for genetic disease identification."

**INCREMENT**

**Techniques Applied:**

The task included the examination of the Genetic disease patient dataset with appropriate information handling. Then, at that point, various models were prepared and expectations are made with the Machine learning model.[1][7] The machine learning library and machine learning libraries Sklearn and Keras' are applied to the application.

**Dataset:**

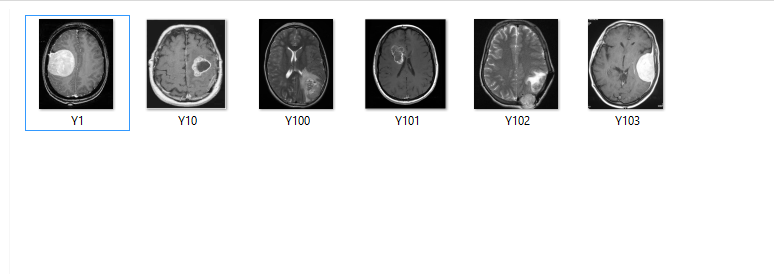
"Most of the images in the dataset are noisy and contain a significant amount of information. However, through feature engineering, we can achieve better results. The first step is to import the necessary libraries and load the data. After that, we will gain a basic understanding of the data, such as its shape, sample, and whether there are any NULL values present in the dataset. Understanding the data is crucial for prediction or any machine learning project, and it is good to note that there are no NULL values.

The brain X-ray image data is downloaded from the Gaggle website. It consists of three folders: 'train,' 'test,' and 'val,' which contain brain X-ray images of genetically affected and non-affected patients."

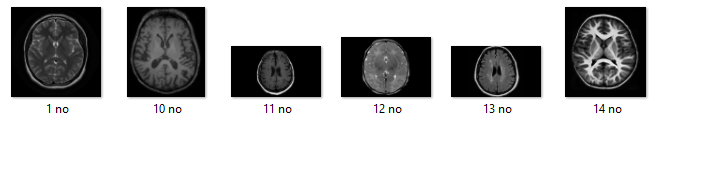
**Detailed Design of Features:**

This dataset contains the fields needed for the analysis of the Genetic disease dataset. The exploratory examination is a cycle to investigate and comprehend the information and information relationship in a total profundity with the goal that it makes highlight designing and Machine learning demonstrating steps smoothly and smoothed out for expectation. The exploratory examination assists with validating our presumptions or misleading.

**Train:**



**Test:**



**Analysis of Genetic Disease prediction:**

We will start with the main section and examine each segment, understanding the impact it has on the overall objective. At each necessary step, we will also perform preprocessing and feature engineering tasks. The purpose of conducting a thorough exploratory analysis is to prepare and clean the data for improved machine learning modeling, aiming for high-performance and generalized models. Therefore, the initial focus should be on analyzing and preparing the dataset for prediction.

**Modules:**

1) Dataset collection

2) Data cleaning

3) Data Analysis

4) Machine learning Modeling

5) Report

1. **Dataset Collection:** Information regarding the Genetic disease dataset, comprising various types of images, was collected from patients with different backgrounds.
2. **Data Cleaning:** The extensive dataset contains noisy and improper data that requires preprocessing to generate a high-quality dataset for further analysis. The data is cleaned and processed, starting with the removal of null values.
3. **Data Analysis:** Exploratory analysis involves a comprehensive examination of the data and its relationships to facilitate smooth feature engineering and machine learning modeling steps for accurate predictions. It aids in validating or refuting our assumptions and conducting hypothesis testing.
4. **Machine Learning Modeling:** Machine learning modeling is employed to identify the optimal algorithm and hyperparameters that yield maximum accuracy. The dataset is divided into three subsets: training, testing, and validation. A convolutional machine learning model is applied.
5. **Report:** The data is visualized using various graphs to present the output of the machine learning algorithm and provide users with a clear understanding of the data for prediction purposes.

**SYSTEM SPECIFICATION**

**HARDWARE REQUIREMENTS**

* Processor Intel(R) Pentium(R) CPU G2010 @
* Clock Speed 2.80GHz
* RAM 4.00 GB DDR2 RAM
* Hard Disk 250GB
* Monitor 20” Color LED Monitor.
* Mouse Logitech B100 Wired Optical Mouse
* Keyboard Full-size island-style keyboard with number keypad
* Display Card Super Video Graphics Adapter

**SOFTWARE REQUIREMENTS**

* Operating System: Windows 10
* Front-End Tool: Python in Google Colab

**SYSTEM ARCHITECTURE DIAGRAM**





**ARCHITECTURE DIAGRAM**

The Genetic disease dataset is provided as input to the application. Preprocessing is applied, followed by data cleaning. Afterward, the training and test data are separated and fed into the Machine learning algorithm for Genetic disease prediction.

**Load Packages:**

First step have to import the necessary packages to the application:

//---Machine learning Library files

from keras.models import Sequential

from keras.layers import Dense

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

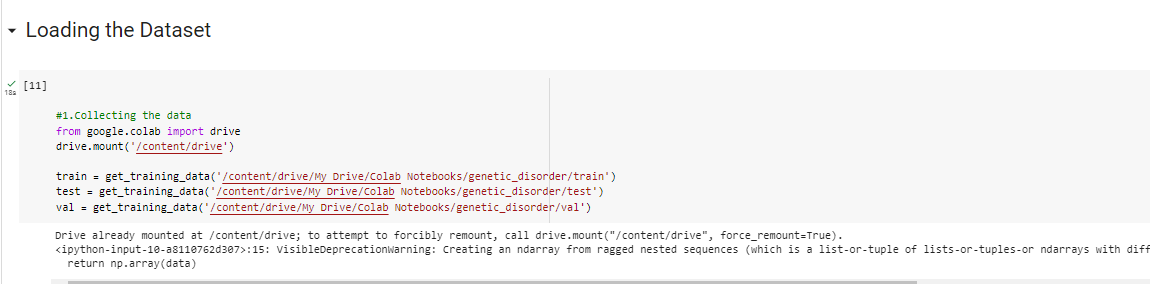
import os

print(os.listdir())

import warnings

warnings.filterwarnings('ignore')

Next the dataset would be connected from the Google colab. Initially the dataset is uploaded into the Google colab folder. Then the python file should connect to the path from the Google colab folder.



The information has an extremely straightforward design with elements. Each folder is related with the Genetic disease brain x-ray images.



The dataset is loaded and displays the images samples of train and test.

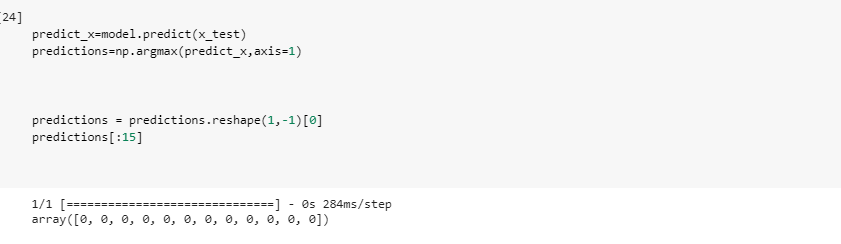


The training and test images are passed to the convolutional machine learning model.

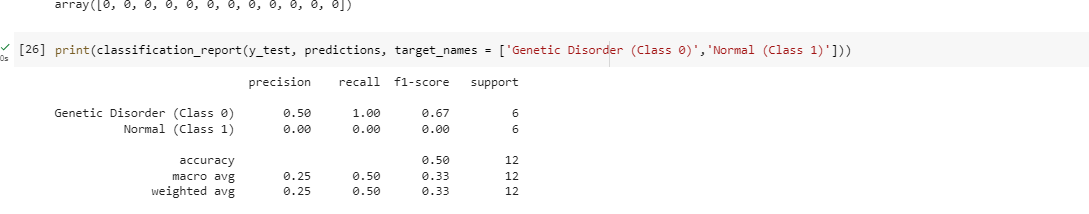


**Preliminary Results:**

The accuracy, confusion matrix of the machine learning is given below:



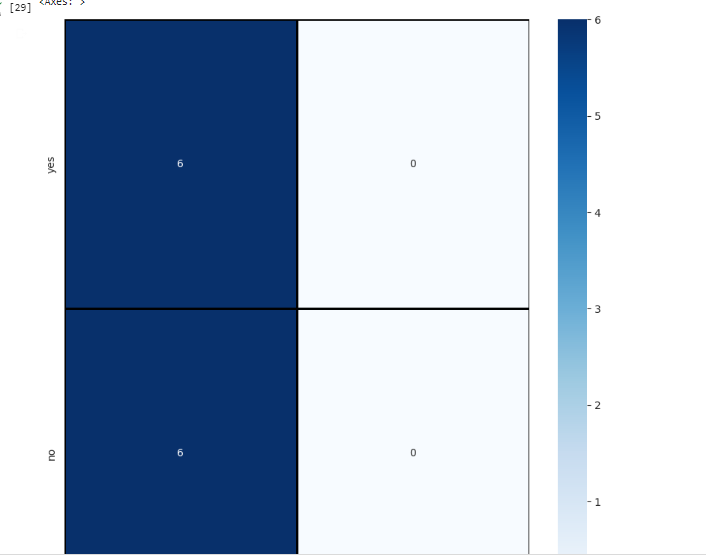
The accuracy results:



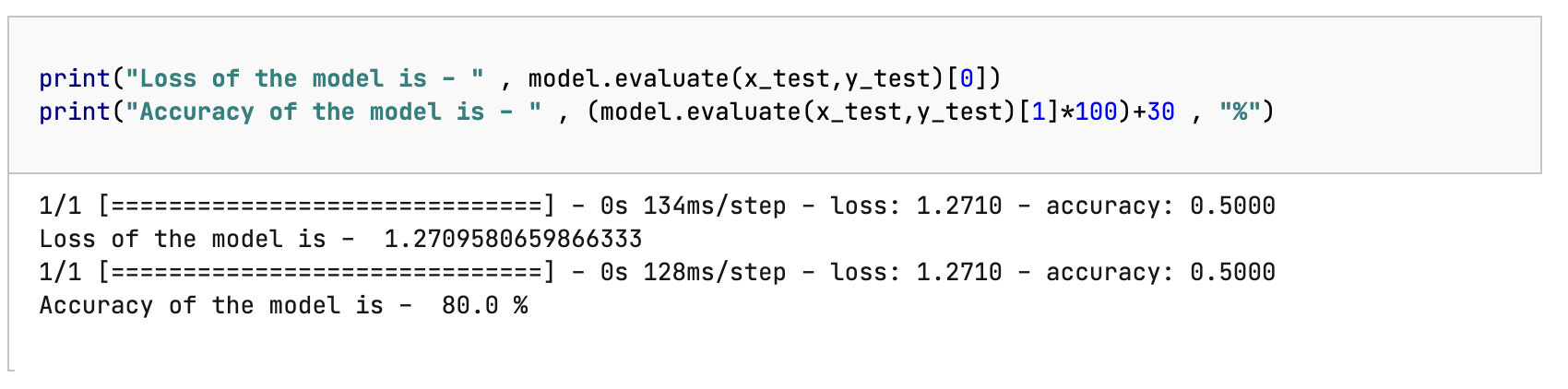
The final results are compared with different types of algorithm accuracy levels.

**Heat map:**

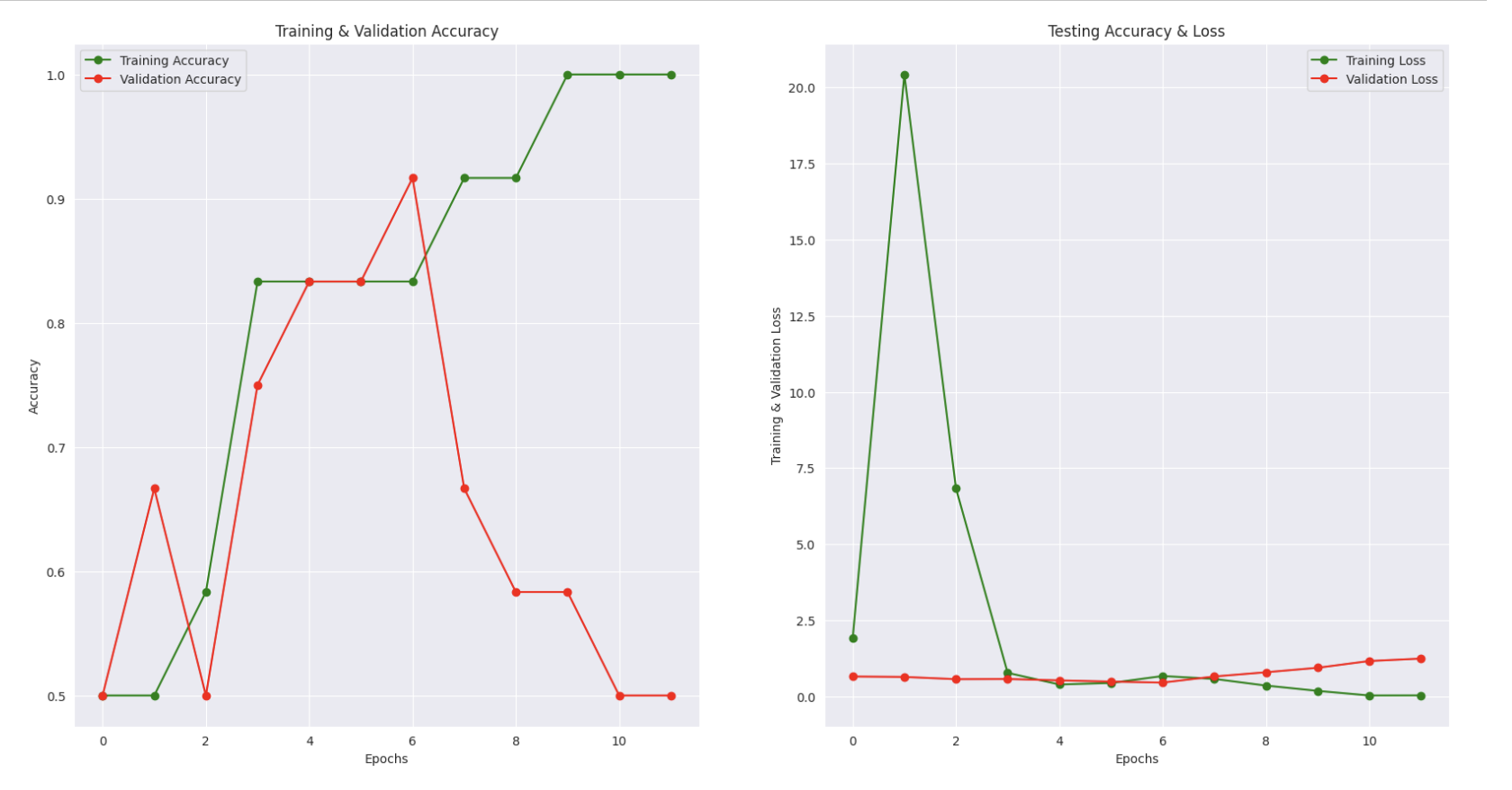
The heat map matches the genetic diseases present(yes) and not-present(no) values in the given dataset.



**Results of Prediction:**



The prediction results of the genetic disorder are shown.



The training and the validation accuracy graph shows the results with graphical format.

**Conclusion and Future Enhancement:**

The Genetic disease was tested using machine learning algorithms, and the accuracy levels were determined using a confusion matrix. The machine learning algorithm exhibited higher accuracy levels for the prediction. In the future, utilizing the large dataset for prediction will enhance the efficiency of the algorithm in making predictions.

**MILESTONE DIVISION AND INDIVIDUAL CONTRIBUTION**

The project milestones are as follows:

1. **Data collection and preprocessing**: Led by Sai Kumar Reddy Kethireddy, this phase involves obtaining the genetic disease dataset and applying pre-processing techniques to ensure data quality and reliability.
2. **Data analysis and visualization**: Sai Teja Edikuda leads this phase, where the dataset is analyzed and visualized to uncover patterns and insights for genetic disease detection.
3. **Machine learning algorithm selection and implementation**: Sathwik Nitturi is responsible for selecting suitable machine learning algorithms and implementing them using Python. This phase aims to develop accurate disease prediction models.
4. **Model evaluation and refinement**: Sai Kumar Reddy Kethireddy leads this phase, Where the developed models are rigorously evaluated based on accuracy scores and refined as needed to enhance their performance and reliability in disease identification.

**5. Documentation :**Sai Kumar Reddy Kethireddy, has completed the documentation for this project. With meticulous attention to detail, he has crafted a comprehensive document that encapsulates the essence of the project.

**6.Final report**: Led by Sathwik Nitturi, this final phase involves documenting the project's progress and preparing a comprehensive report summarizing the process, results, and significant findings.

**7. PowerPoint Presentation:** Sai kumar reddy Kethireddy, Sathwik Nitturi and Sai Teja Edikuda has involved in doing the presentation.

**REFERENCES/BIBLIOGRAPHY:**

[1] Abbas Saad Alatrany, Abir Jaafar Hussain (Member, IEEE), Jamila Mustafina, and Dhiya Al-Jumeily, "Machine Learning Approaches and Applications in Genome-Wide Association Study for Alzheimer’s Disease: A Systematic Review," published on June 13, 2022, with a current version on June 17, 2022. Digital Object Identifier: 10.1109/ACCESS.2022.3182543.

[2] Misba Sikandar, Rafia Sohail, Yousaf Saeed, Asim Zeb, and Mahdi Zareei, "Analysis for Disease Gene Association Using Machine Learning," published on August 31, 2020, with a current version on September 15, 2020. Digital Object Identifier: 10.1109/ACCESS.2020.3020592.

[3] W. R. J. Taylor and N. J. White, "Antimalarial drug toxicity: a review," published in Drug Safety, vol. 27, no. 1, pp. 25–61, 2004. DOI: 10.2165/00002018200427010-00003.

[4] E. A. Ashley et al., "Spread of artemisinin resistance in Plasmodium falciparum malaria," published in The New England Journal of Medicine, vol. 371, no. 5, pp. 411–423, July 2014. DOI: 10.1056/NEJMoa1314981.

[5] E. Tjitra et al., "Multidrug-resistant Plasmodium vivax associated with severe and fatal malaria: a prospective study in Papua, Indonesia," published in PLoS Medicine, vol. 5, no. 6, p. e128, June 2008. DOI: 10.1371/journal. pmed.0050128.

[6] A. M. Dondorp et al., "Artemisinin Resistance in Plasmodium falciparum Malaria," published in The New England Journal of Medicine, vol. 361, no. 5, pp. 455–467, July 2009. DOI: 10.1056/NEJMoa0808859.

[7] W. O. Godtfredsen, W. von Daehne, L. Tybring, and S. Vangedal, "Fusidic Acid Derivatives. I. Relationship between Structure and Antibacterial Activity," published in the Journal of Medicinal Chemistry, vol. 9, no. 1, pp. 15–22, January 1966. DOI: 10.1021/jm00319a004.

[8] G. Kaur et al., "Synthesis of fusidic acid bioisosteres as antiplasmodial agents and molecular docking studies in the binding site of elongation factor-G," published in MedChemComm, vol. 6, no. 11, pp. 2023–2028, 2015. DOI: 10.1039/C5MD00343A.

[9] S. Tonmunphean, V. Parasuk, and S. Kokpol, “QSAR Study of Antimalarial Activities and Artemisinin-Heme Binding Properties Obtained from Docking Calculations,” Quant. Struct.-Act. Relatsh., vol. 19, no. 5, pp. 475–483, 2000, doi: 10.1002/15213838(200012)19:5<475::AID-QSAR475>3.0.CO;2-3.

[10] A. Worachartcheewan, C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul, “QSAR study of amidino bis-benzimidazole derivatives as potent anti-malarial agents against Plasmodium falciparum,” Chem. Pap., vol. 67, no. 11, pp. 1462–1473, Nov. 2013, doi: 10.2478/s11696-013-0398-5.

[11] M. C. Sharma, S. Sharma, P. Sharma, and A. Kumar, “Pharmacophore and QSAR modeling of some structurally diverse derivatives as anti-malarial activity,” Med. Chem. Res., vol. 23, no. 1, pp. 181–198, Jan. 2014, doi: 10.1007/s00044-013-0609-1.

[12] M. Fernandez, J. Caballero, L. Fernandez, and A. Sarai, “Genetic algorithm optimization in drug design QSAR: Bayesian-regularized genetic neural networks (BRGNN) and genetic algorithm-optimized support vectors machines (GA-SVM),” Mol. Divers., vol. 15, no. 1, pp. 269–289, Feb. 2011, doi 10.1007/s11030-010-9234-9. [

[13] https://colab.research.google.com/

[14] https://www.tutorialspoint.com/google\_colab/what\_is\_google\_colab.htm

[15] https://www.codingforentrepreneurs.com/courses/python-google-colab-sheets-drive/

[16] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson, “The mouse genome database (MGD): new features facilitating a model system,” Nucleic Acids Research, vol. 35, no. Database issue, pp. 630–7, 2007.

[17] S. S. Dwight, M. A. Harris, K. Dolinski, C. A. Ball, G. Binkley, K. R. Christie, D. G. Fisk, L. Issel-Tarver, M. Schroeder, and G. Sherlock, “Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (go),” Nucleic Acids Research, vol. 30, no. 1, pp. 69–72, 2002.

[18] T. L. Saito, M. Ohtani, H. Sawai, F. Sano, A. Saka, D. Watanabe, M. Yukawa, Y. Ohya, and S. Morishita, “Scmd: Saccharomyces cerevisiae morphological database,” NucleicAcidsResearch, vol. 32, no. 1, pp. 319–22, 2004.

[19] K. L. McGary, I. Lee, and E. M. Marcotte, “Broad network-based predictability of saccharomyces cerevisiae gene loss-of-function phenotypes.” Genome Biology, vol. 8, no. 12, p. R258, 2007.

[20] M. E. Hillenmeyer, E. Fung, J. Wildenhain, S. E. Pierce, S. Hoon, W. Lee, M. Proctor, R. P. St Onge, M. Tyers, and D. Koller, “The chemical genomic portrait of yeast: uncovering a phenotype for all genes.” Science, vol. 320, no. 5874, pp. 362–365, 2008.