**MACHINE LEARNING PROJECT – CRN 30460**

**PROJECT: 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 |

**ABSTRACT**

Identifying disease genes in the human genome is crucial in biomedical research. However, the limited number of confirmed disease genes and genetic heterogeneity pose challenges to discovering new disease genes using machine learning (Alatrany et al., 2022). Recent approaches employ 'guilty by association' to explore the link between disease phenotypes and causative genes (Sikandar et al., 2020). Imbalanced data issues in disease gene identification necessitate semi-supervised techniques like label propagation and positive-unlabeled learning. Additionally, ensemble learning models integrate multiple biological sources to enhance predictive performance. This thesis presents three computational models for identifying candidate disease genes.

**PROJECT EXECUTION PLAN**

This work utilizes a genetic disease dataset and applies preprocessing techniques to remove noisy and null value data. Subsequently, we analyze and visualize the data for further processing. We select machine learning algorithms for disease prediction.

**CONTRIBUTION**

"Genetic Disease Detection with Machine Learning" is a Python application that employs patient record attributes to identify genetic diseases. It contributes significantly to genetic disease research by providing a reliable and accurate tool for disease identification.

**EVALUATION**

The evaluation of the project relies on the prediction results obtained from the machine learning algorithm, with the accuracy score playing a crucial role in assessing the dataset.

The application is developed using Google Colab, a Python tool that enables direct execution on any computer system with an internet connection, eliminating the need for specific software installations. With Colab, developers can create and run the application on a cloud server, which already has the required Python library files. The Colab environment includes machine learning algorithm libraries, enabling the utilization of deep learning algorithms for genetic disease detection.

The dataset is divided into training, testing, and validation sets, and it comprises genetic disease data.

**MILESTONE DIVISION AND INDIVIDUAL CONTRIBUTION**

The project milestones are as follows:

1. **Data collection and pre-processing**: 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 and 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.

**Empirical Evaluation:**

We will be using various supervised machine learning algorithms such as:

● Random Forest Classifier

● Logistic Regression

● K-Nearest Neighbour Classifier

**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.