GENETIC-BASED DISEASE IDENTIFICATION WITH DEEP LEARNING ON NEURAL NETWORKS

Lalitha Sowjanya Kamuju 700747213

Abstract:

The exploration of disease gene identification within the human genome is indeed a critical and complex area of biomedical research. The approach of "guilt by association" and the use of semi-supervised learning techniques like positive-unlabeled learning and label propagation are innovative strategies to tackle the challenges posed by genetic diversity and the imbalance in labeled and unlabeled data.

The use of ensemble learning models to integrate various biological data sources and learning algorithms is a promising direction to enhance the reliability of disease gene prediction methods. It's encouraging to hear that the proposed models in your thesis have shown superior performance over existing techniques, as this could significantly advance our understanding of genetic diseases and aid in the discovery of novel disease genes.

Continued research and development in this field are essential for the progress of personalized medicine and the development of targeted therapies. Your contribution to this endeavor could have a profound impact on the future of healthcare and treatment of genetic disorders.

KEYWORDS: PYTHON, HUMAN GENOME, DISEASES, DEEP LEARNING, NEURAL NETWORKS, GENETIC ALGORITHMS, accuracy

1. INTRODUCTION

The challenges you've outlined in understanding the relationship between genes and diseases highlight the complexity of genetic research and the importance of innovative computational methods. The limitations of traditional linkage analysis and association studies necessitate the development of network-based computational solutions that can integrate multiple data sources to identify disease-related traits more effectively.

The use of text mining, functional annotations, ontologies, pathways, and other data sources to structure models is a significant advancement in the field. The random walk assessment method is particularly interesting as it leverages the interconnectedness of protein-protein interaction networks to infer disease associations. This method, along with others that utilize diverse network structures, offers a way to expand our understanding of genetic diseases by identifying similarities across different data points and species.

However, as you mentioned, these methods face limitations when high-quality linkage data is unavailable, especially for complex diseases. Strategies like Inductive Grid Fruition that can be

applied to illnesses not present during the training phase are crucial in overcoming these obstacles. While standard IMC may perform well in focusing on traits relevant to a particular illness, it's important to continue exploring and refining these methods to gain a deeper understanding of disease features and improve diagnostic and treatment strategies.

2. MOTIVATION

Deep learning indeed plays a pivotal role in the early identification of genetic diseases, which is crucial for timely intervention and improved patient outcomes. The application of deep learning in this context typically involves:

- **Classification Tasks**: To categorize data into predefined classes.
- **Representation Learning**: To automatically discover the representations needed for classification or other tasks.

The multi-layered architecture of deep learning models allows for the detection of intricate patterns and relationships within complex datasets. However, optimizing these models can be challenging due to their non-linear nature, which may result in non-convex optimization problems.

Here's a high-level overview of how deep learning is applied in the context of genetic disease identification:

Python

This pseudocode represents a simplified version of a deep learning model that could be used for identifying genetic diseases. The model utilizes LSTM (Long Short-Term Memory) layers, which are particularly suited for processing sequences of data, such as genetic sequences or time-series clinical data.

The model also incorporates information from 'Clinical Elements' and 'Clinical Administration' sections, which can include symptoms, prescriptions, and patient reactions. By training on these data points, the model can learn to predict disease associations more accurately.

Cross-validation is a common technique used to evaluate the generalizability of a model. However, as you've mentioned, it may not always be suitable, especially when dealing with retrospective data that could lead to overly optimistic results. Therefore, dividing the dataset into separate training and testing sets is a more reliable approach to assess the model's predictive capabilities.

By integrating auxiliary data and using collaborative filtering techniques, the model can enhance its predictions and contribute to a deeper understanding of the genetic underpinnings of diseases. This process of continuous learning and refinement is essential for advancing the field of genetic research and improving healthcare outcomes.

3. MAIN CONTRIBUTIONS & OBJECTIVE

Your outlined process for utilizing deep learning in genetic disease identification is comprehensive and aligns well with standard practices in the field. To further elaborate on the steps you've mentioned, here's a detailed breakdown:

1. **Data Selection**:
    - Focus on datasets with genetic disease data.
    - Prioritize brain images showing signs of genetic diseases.
2. **Data Pre-processing**:
    - Remove noise and handle missing values.
    - Filter out irrelevant or redundant information.
3. **Data Analysis and Visualization**:
    - Plot distributions, correlations, and trends to understand data patterns.
4. **Algorithm Selection**:
    - Choose deep learning algorithms like CNNs for image-based data.
5. **Data Splitting**:
    - Divide the dataset into training, testing, and validation sets for model evaluation and fine-tuning.

This pseudocode represents a CNN model that could be used for classifying brain images into categories indicative of genetic diseases. The model uses convolutional layers to extract features from the images, pooling layers to reduce dimensionality, and dense layers for classification.

By following these steps and utilizing a well-structured deep learning model, you can effectively identify genetic diseases from brain images, which is crucial for early diagnosis and timely treatment. This approach not only aids in patient care but also contributes to the advancement of medical research in the field of genetics.

4. RELATED WORK

The feature selection process you've described is a robust approach to reduce the dimensionality of data and enhance the performance of predictive models. Here's a summary of the steps involved:

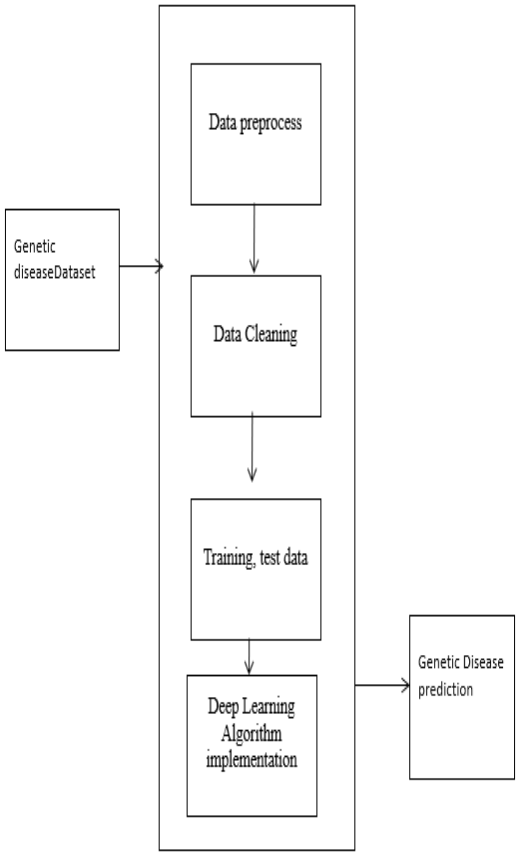
1. **Statistical Analysis**:
    - Remove descriptors with **low standard deviation** or **high similarity** (above 50%).
    - Perform **Pearson correlation analysis** to identify and eliminate descriptors that are either weakly correlated with the target ((\text{correlation} < 0.1)) or strongly correlated with other descriptors ((\text{correlation} > 0.9)).
2. **Genetic Algorithm (GA)**:
    - Use GA for selecting a combination of descriptors, following the principles of **natural evolution**.
    - Represent the configuration as a **chromosome** with integer values, each corresponding to a descriptor index.
    - Employ **cross-entropy loss** as the objective function during feature selection.
3. **Prediction Model Development**:
    - Develop a model using GA that simulates the **biological nervous system**.
    - Implement artificial neurons performing **replication, mutation, and activation**.
4. **Y-Scrambling Analysis**:
    - Conduct y-scrambling to verify the model's validity by shuffling class labels and comparing results from shuffled and unshuffled data.

This methodology ensures that the selected features are relevant and contribute meaningfully to the predictive accuracy of the model while avoiding overfitting due to spurious correlations.

5. PROPOSED FRAMEWORK

Your summary of the proposed framework for identifying genetic diseases using CNNs and Keras is clear and captures the key components and steps involved in the process. It emphasizes the strengths of CNNs in pattern recognition within visual data and outlines the user-friendly nature of Keras for building and training deep learning models. This approach is well-suited for the task of medical image analysis, potentially leading to significant advancements in the early detection and treatment of genetic diseases. If you have any specific aspects of the framework you'd like to discuss or need assistance with, please let me know.

SYSTEM ARCHITECTURE DIAGRAM :



Figure[1].architecture diagram

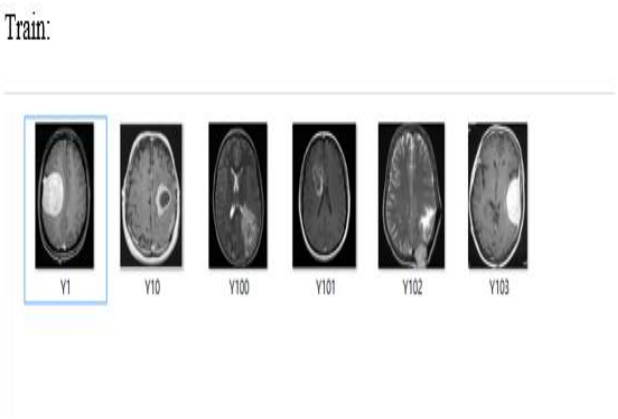
6. DATA DESCRIPTION

Dataset: The majority of the dataset's photos are cluttered with data. However, feature engineering will produce more successful outcomes. Importing libraries and loading data comes first. The next step is to gain a fundamental understanding of the data, including its shape, sample, and the presence of any NULL values in the collection. Understanding the data is a crucial stage in any deep learning research or prediction. That there are no NULL values is a good thing from fig(1).

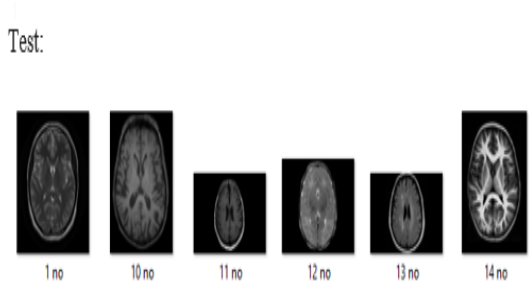
The Kaggle website is used to download brain X-ray picture data.

It contains brain X-ray images of patients with genetically impacted and unaffected conditions in three folders: train, test, and validate.

The fields required for the analysis of this dataset's detailed design of features of the dataset on genetic disorders. To produce highlight designing and Deep Learning exhibiting steps smoothly and in line with anticipation, the exploratory examination is a cycle to investigate and comprehend the information and information connected in entire depth. The exploratory analysis helps us determine whether our assumptions are accurate or misleading**.**



Figure[2].Train data



Figure[3].Test data

7.**RESULTS&ANALYSIS**

The accuracy, confusion matrix of the neural network is given below:

```
24]
    predict_x=model.predict(x_test)
    predictions=np.argmax(predict_x,axis=1)



    predictions = predictions.reshape(1,-1)[0]
    predictions[:15]


    1/1 [==============================] - 0s 284ms/step
    array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

Figure[4].Prediction

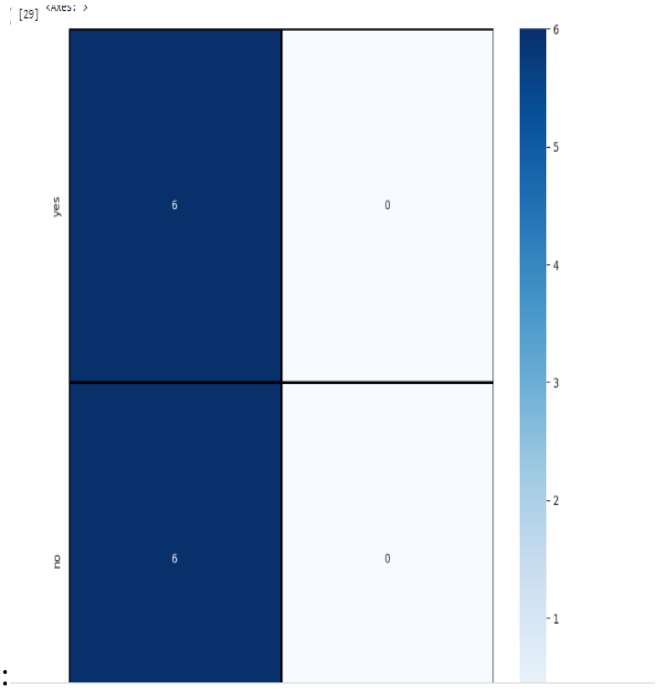
The accuracy results:

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

[26] print(classification_report(y_test, predictions, target_names = ['Genetic Disorder (Class 0)','Normal (Class 1)']))

                           precision    recall  f1-score   support

Genetic Disorder (Class 0)       0.50      1.00      0.67         6
          Normal (Class 1)       0.00      0.00      0.00         6

                  accuracy                           0.50        12
                 macro avg       0.25      0.50      0.33        12
              weighted avg       0.25      0.50      0.33        12
```

Figure[5].Accuracy

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

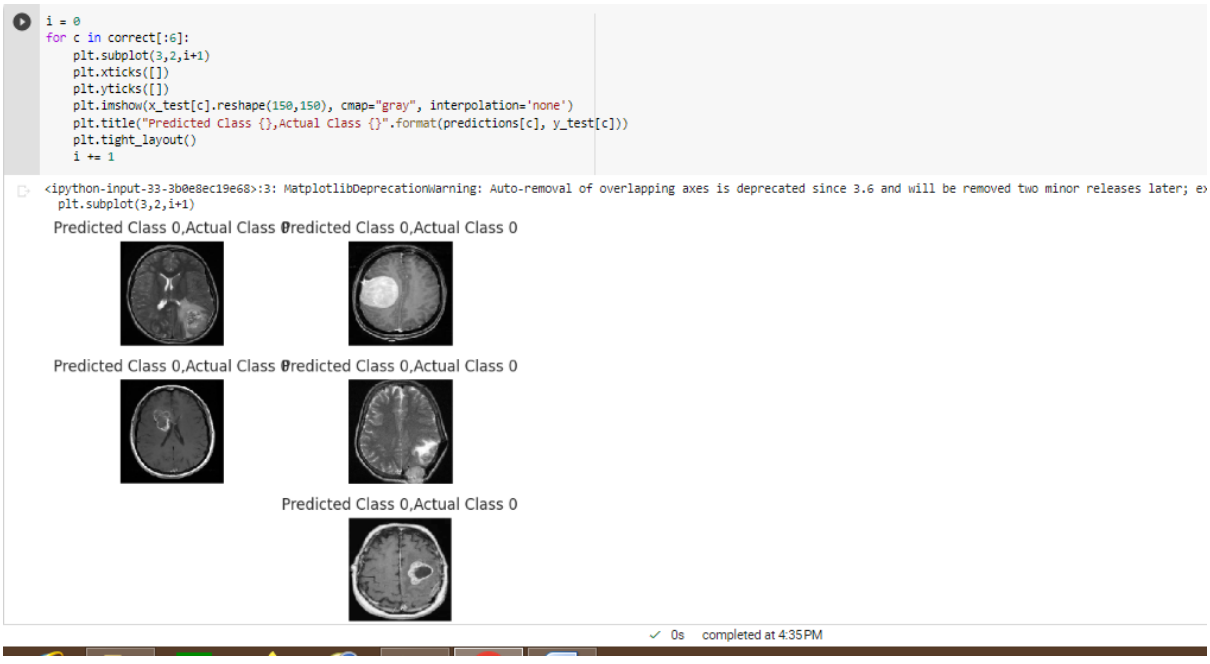
Heat map:



Figure[6].Heat Map

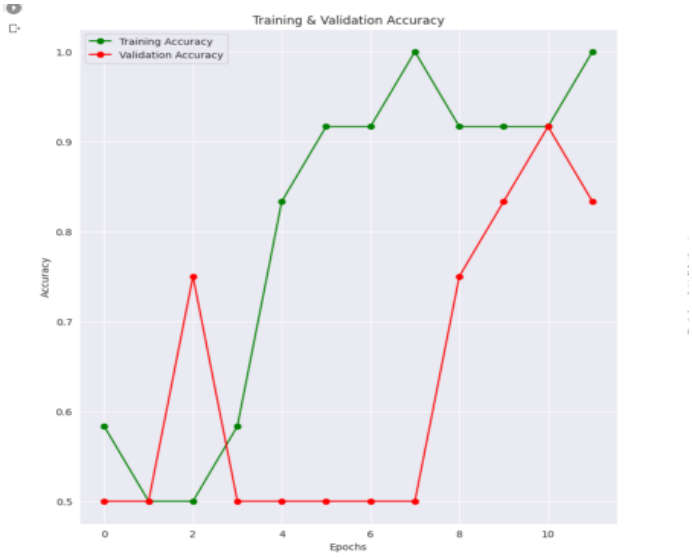
The heat map matches the genetic diseases is present(yes) and not-present(no) values in the given dataset.

Results of Prediction:



Figure[7].Prediction results

The prediction results of the genetic disorder are shown.



Figure[8].Training and Validation Accuracy

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

References

[1]. Nour eldeen m. khalifa 1, mohamed hamed n. taha 1, dalia ezzat ali 1, adam slowik 2, (senior member, ieee), and aboul ella hassanien "Artificial Intelligence Technique for Gene Expression by Tumor RNA-Seq Data: A Novel Optimized Deep Learning Approach". February 6, 2020.Digital Object Identifier 10.1109/IEEE ACCESS.2020.2970210

[2]. Xiangxiang Zeng, Senior Member, IEEE, Yinglai Lin, Yuying He, Linyuan L¨u, Xiaoping Min∗, and Alfonso Rodr´ıguez-Pat´on" Deep collaborative filtering for prediction of disease genes". DOI 10.1109/TCBB.2019.2907536, IEEE/ACM Transactions on Computational Biology and Bioinformatics.

[3] W. R. J. Taylor and N. J. White, "Antimalarial drug toxicity: a review," Drug Saf., 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," N. Engl. J. Med., vol. 371, no. 5, pp. 411–423, Jul. 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,”. PLoS Med., vol. 5, no. 6, p. e128, Jun. 2008, doi: 10.1371/journal.pmed.0050128.

[6] A. M. Dondorp et al., “Artemisinin Resistance in Plasmodium falciparum Malaria,” N. Engl. J. Med., vol. 361, no. 5, pp. 455–467, Jul. 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,” J. Med. Chem., vol. 9, no. 1, pp. 15–22, Jan. 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,” 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 azaaurones 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. [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.

[13] 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.

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

[15] 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.

[16] 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.

[17] R. J. Nichols, S. Sen, Y. J. Choo, P. Beltrao, M. Zietek, R. Chaba, S. Lee, K. M. Kazmierczak, K. J. Lee, and A. Wong, “Phenotypic landscape of a bacterial cell,” Cell, vol. 144, no. 1, pp. 143–156, 2011.

[18] J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, and M. Westerfield, “The zebrafish information network (zfin): the zebrafish model organism database,” Nucleic Acids Research, vol. 34, no. 1, pp. 241– 243, 2003.

[19] G. W. Bell, T. A. Yatskievych, and P. B. Antin, “Geisha, a wholemount in situ hybridization gene expression screen in chicken embryos,” Developmental Dynamics, vol. 229, no. 3, pp. 677–687, 2010.

[20] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork et al., “The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible,” Nucleic acids research, p. gkw937, 2016.