Software Proposal Document for project Genetics

Kareem Ehab, Mohamed Moataz, Youssif Assem, Ahmed Gamal

supervised by:DR. Fatma Helmy and Eng. Ahmed Hazem

November 8, 2021

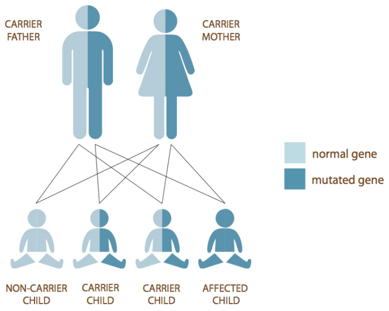
Proposal Version	Date	Reason for Change
1.0	24-October-2021	Proposal First version

Table 1: Document version history

GitHub: https://github.com/YoussifAssem/Genetics

Abstract

In humans, genes are the essential factor. A sickness induced in part or wholly by a departure from the usual DNA sequence is known as a hereditary ailment. Anyone could carry altered genes and pass them on to his children without even realizing it. Consider the case where we have a model that can predict which genetic disease will emerge in future generations from studying mutations of both genders. A prediction will occur to the new generation of that specific genetic disease based on what was learned from those mutations after taking the gene. What follows is a detailed discussion of what and how the model will work and what disease we will focus on.



1 Introduction

1.1 Background

Artificial intelligence is increasingly being utilised to target and prevent diseases.[14] Artificial intelligence has being rapidly adopted to assist clinicians in diagnosis, illness tracking, prevention, and control in order to meet these problems. A lot of research is being done in this field by applying data mining and machine learning techniques to assist medical professionals. If such procedures work for certain diseases, they should work for genetic disorders as well. A single gene mutation, numerous gene mutations , a combination of gene mutations and environmental variables, or chromosome damage can all be causes of genetic illnesses (changes in the number or structure of entire chromosomes, the structures that carry genes). Now it is much easier to benefit from the use of artificial intelligence to detect and prevent these genetic diseases before it is too late.

1.2 Motivation

1.2.1 Academic

People always suffered from genetic disorders, The first genetic disorder mapped using DNA polymorphisms. was in 1983 [1] People always suffered from genetic disorders, The first genetic disorder mapped using DNA polymorphisms. was in 1983 . When a gene mutation occurs, the nucleotide is in the wrong order, which means the coded instructions are incorrect, resulting in the production of defective proteins or the alteration of control switches. Gene mutations occur throughout life and can cause the body to malfunction. They can occur as a result of copying errors produced during cell division and replication. Viruses, exposure to radiation (such as the sun), and chemicals can also cause them (such as smoking). In some genetic diseases, Mutations can be passed down down the generations from one or both parents. They can be found in both the egg and sperm cells. We all have certain inherited gene mutations that have no negative consequences. Others make certain people more susceptible to certain illnesses. Researchers have identified three main causal groups for genetic diseases: single-gene errors, multiple-gene alterations, and chromosomal abnormalities. While many hereditary disorders present at birth, others do not display symptoms until later in life, anywhere from 30 to 70 years. According to the article, the problem is partially solved. [7], XGBoost is a model built on Boosting Tree models.

1.2.2 Business

Defining a business requirement is an important part of the enterprise analysis process. Understanding and determining the system's aim, as well as expressing a strategic direction, are all part of this process. Furthermore, we shall document any critical concerns about the project's success, issues, or challenges. our application aligns with the medical business goals and objectives of the market, it will target a new segment of patients, as this type of blood analysis has not been found before in the market, and in turn, it will increase the number of arrivals at the laboratories,our team is investigating the business challenge and opportunity to ensure that there is a compelling rationale to continue forward and meet market demands We'll make sure the software helps to improve and provide value to the company.

1.3 Problem Statement

Many issues arise as a result of hereditary illnesses. For example, ulcerative colitis disease develops when a person carries the Mediterranean fever gene. This condition is known as concurrent illness. We aim to build a model that will take two parents' defective genes and forecast what genetic disorders will occur in the next generation.

2 Project Description

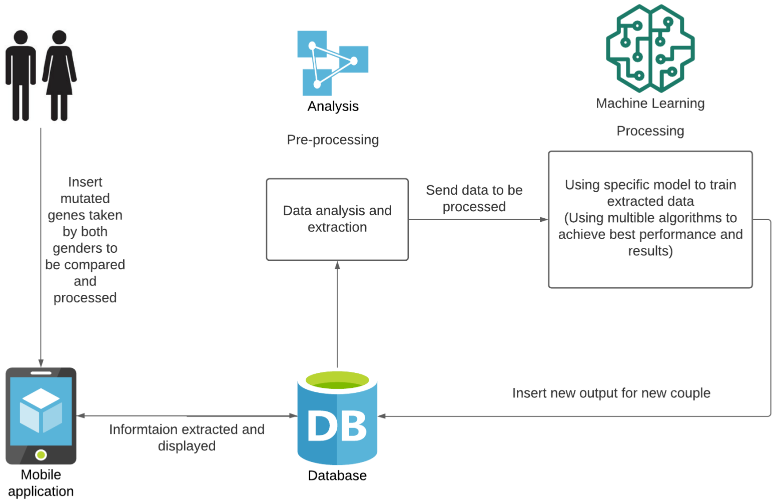
2.1 Objectives

Our main objective is to create a system model that can detect the genetic disorder in the new generation resulting from two new couples based on reading the mutated genes from both genders' SNPs in a certain genetic disease. The model will inform them what percentage of genetic illnesses their children will inherit. It will be extremely beneficial for people with historic genetic disorders in their family trees to acknowledge that there could be a problem and take steps to prevent it from being a problem in the future.

2.2 Scope

In the proposed system we will be able to predict the genetic disease by taking the data of DNA analysis of both father and mother (input), where the model compares between these data analysis and predicts the upcoming disease in the child (output)

2.3 Project Overview


Analysis
Pre-processing
Machine Learning
Processing
Insert mutated genes taken by both genders to be compared and processed
Data analysis and extraction
Send data to be processed
Using specific model to train extracted data (Using multible algorithms to achieve best performance and results)
Insert new output for new couple
Informtaion extracted and displayed
DB
Mobile application
Database

Our model will predict the genetic disorders in the new generation of a new couple based on their mutated genes. We are presenting a system process that will take the mutated genes given by the couple. Then It will be taken to the pre-processing phase where there It will extract the targeted data that will then be processed. After that the extracted data will be fed to the machine learning model which uses multiple algorithms to achieve and maximize performance and results as it will be compared with the data in the original data set. Finally, the information in the end will presented and displayed in a user-friendly mobile application.

2.4 Stakeholder

2.4.1 Internal

Memeber	Job
Youssif Assem (team leader)	back-end developer, research papers, document writing and front-end
Kareem Ehab	back-end developer, research papers, document writing and front-end
mohamed moataz	back-end developer, research papers, document writing and front-end
Ahmed gamal	back-end developer, research papers, document writing and front-end

2.4.2 External

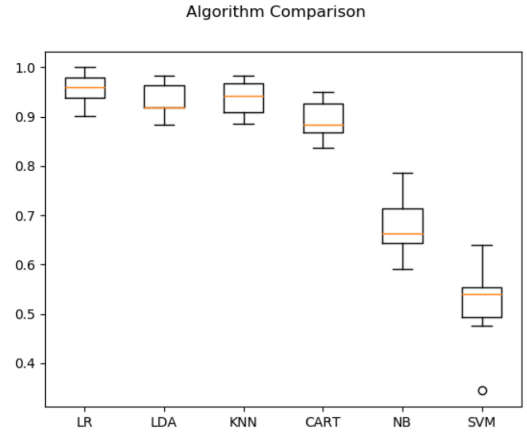
the external stakeholders will be people in general

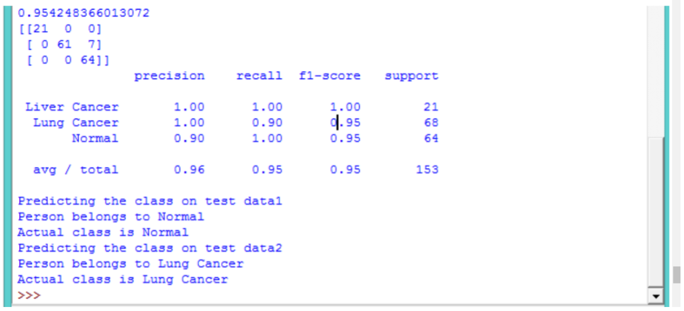
3 Similar System

3.1 Academic

The goal of this study was to see if machine learning techniques and the polygenic risk score might be used to predict hereditary illnesses. They used a range of machine learning approaches, including support vector machines, Random forests, and k-nearest neighbours, all of which have shown to be effective in forecasting the risk of complicated diseases based on clinical data. For example, random forest, AdaBoost, XGBoost, support vector machines,neural network and superlearner are utilised without models in Gao's study [12]of parkinson disease. It's important to understand that the polygenic risk score adds the is based on a predetermined model that will add the the set of risk alleles to a specific complex disease. Based on their study the use of genetic disease type and PRS being examined will determine how much machine learning techniques are used, for example, for psychological disorders the PRS is generated by a program called PRSice and genome-wide association studies as a complement to polygenic risk prediction. Other techniques are used such as multivariate relevance vector regression where RVR is a a probabilistic pattern recognition approach based on nuclei.

F. Francis and T. N. Namitha [4]proposed a model for to monitor tens of thousands of genes and their term levels concurrently. The fundamental purpose of gene expression microarray data pre-processing is to find a limited number of genes that may be exploited to improve classification accuracy and efficiency from a very high-dimensional gene expression dataset. Gene selection is an important step in the pre-processing of microarray data. By removing unnecessary genes from microarray data, an effective gene selection strategy improves classification accuracy. It also reduces the size of the microarray data, which speeds up the classification process.
They classify the features of the genes with different classification methods then compare the algorithm's accuracy afterwards. Logical Regression, Linear discriminant analysis, KNearest neighbor, classification and regression trees, naive bayes and support vector machine ratios are provided in the figure below Different classification methods are given the feature genes obtained by the SVM-RCE stage. Then compare the algorithms' categorization accuracy. The findings of the logical Regression(LR), Linear Discriminant Analysis(LDA), KNearest Neighbor(KNN), Classification Regression Trees(CART), Naive Bayes(NB), and Support Vector Machine(SVM) comparisons are provided in the figure below





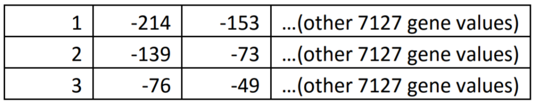
```
0.954248366013072
[[21  0  0]
 [ 0 61  7]
 [ 0  0 64]]
               precision    recall  f1-score   support

 Liver Cancer       1.00      1.00      1.00        21
  Lung Cancer       1.00      0.90      0.95        68
       Normal       0.90      1.00      0.95        64

  avg / total       0.96      0.95      0.95       153

Predicting the class on test data1
Person belongs to Normal
Actual class is Normal
Predicting the class on test data2
Person belongs to Lung Cancer
Actual class is Lung Cancer
>>>
```

Then summarized the results

U. K. Dey and M. S. Islam [6] To appropriately diagnose the type of leukaemia that a certain person has multiple algorithms were utilised. They used a dataset from kaggle that contained the genetic expression of 72 persons, each of whom had 7129 genes. Samples of before and after pre-processing phase and also results after prediction from Random Forest and XGBoost algorithms

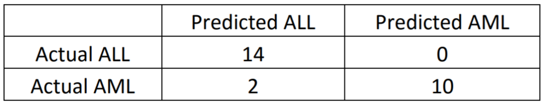
Gene Description	Gene Accession Number	1	call	...

before pre-processing



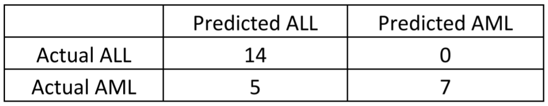
1	-214	-153	…(other 7127 gene values)
2	-139	-73	…(other 7127 gene values)
3	-76	-49	…(other 7127 gene values)

after pre-processing



	Predicted ALL	Predicted AML
Actual ALL	14	0
Actual AML	2	10

XGBoost results



	Predicted ALL	Predicted AML
Actual ALL	14	0
Actual AML	5	7

Random Forest result

Python was the only tool utilised in this study and different packages were used like numpy, pandas and others. The computations were carried out using the test set, and the results were compared to the real data. If the forecast was 0, the person was assumed to have ALL, and if it was 1, they had AML. Using these results on a confusion matrix and calculating all true and false aml values with their precision, the final scores were calculated.

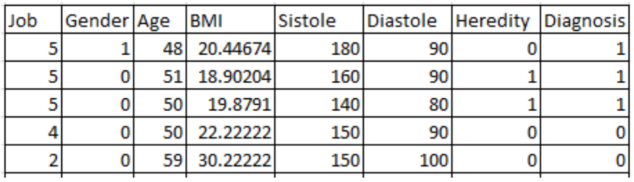
[15] This experiment was conducted using the GEO database, which stores curated gene expression of various illnesses and is maintained by the National Center for Biotechnology Information (NCBI) (Edgar et al. 2002). An Agilent 44 K Human whole-genome array G4112F, GPL6480 (Kuwano et al. 2011) microarray was used to extract oligonucleotide gene expressions from blood samples.

GSE26415 is the accession number for the datasets used in this investigation. There are two separate categories in the dataset: autistic and non-autistic. Each category contains 21 samples. There are 42 samples in the collection, totaling 19,194 gene expression probes in total

As Alzheimer's disease is caused by environmental factors that affect the human brain, Ronghui Ju et.al [13] suggested using deep learning in conjunction with the brain network and clinically significant information such as age, ApoE gene, and gender of the subjects for an earlier examination of the disease. A neural network was established by estimating functional connections in the brain region using resting-state functional magnetic resonance imaging (R-fMRI) data. They did this to grant a described data of the initial Alzheimer's. They also used a deep network like autoencoder to help them.

Manasee Kurkure and Anuradha Thakare presented a lung cancer detector based on the effective use of the best characteristics of Genetic Algorithm and Naive Bays Classifier [3]. They had a dataset consisting of lung cancer images from various people. The classifier would take these photos and classify them as cancerous or non-cancerous, with the entire process being improved using a genetic algorithm.

A research study by Mira Kania Sabariah, MT,Aini Hanifa ST, Siti Sa'adah, MT [2] aimed to create a system to detect Diabetes Mellitus (DM) type II using multiple methods such: Random Forest (RF), Classification and Regression Tree (CART) by applying a dataset taken from public health care shown in the figure below



Job	Gender	Age	BMI	Sistole	Diastole	Heredity	Diagnosis
5	1	48	20.44674	180	90	0	1
5	0	51	18.90204	160	90	1	1
5	0	50	19.8791	140	80	1	1
4	0	50	22.22222	150	90	0	0
2	0	59	30.22222	150	100	0	0

dataset used

Dina Y. Mikhail [8] presented a technique that uses the TP53 gene mutation to identify pre-cancer. Two ways have been used to apply the methodology as described. The first method predicts whether or not a person carries cancer-causing mutations. The second technique, which identifies mutations, is separated from the first to see if the patient's gene mutation is linked to a specific illness (cancer), such as lung cancer, head and neck cancer, breast cancer, and so on.

Joseph M. De Guia, Dr. Madhavi Devaraj Dr. Larry A. Vea [5] proposed a model that classifies cancer using gene expression data by feature selection. They used a dataset containing leukemia which had two classes: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). There are 72 samples in the dataset, including 7129 gene expression instances. The training data contains 38 samples, with 27 belonging to the ALL cancer class and 11 to the AML cancer class. There are 34 samples in the test data: 20 samples from the ALL class and 14 samples from the AML class. To reduce the number of genes in the training model, the normalisation procedure used calculated for T-test mean difference values.

Yifan Gao, Minhan Guo, Haoran Wang Yajin Li [10]has proposed a similar system but for grastic cancer that uses an adaptive learning pipline. It consists of an automatic feature selection component, differential gene screening, machine learning methods and the grid search method to find the best parameters and methods of fitting the model. The dataset they gathered will go through two different pipelines after they are done with data pre-processing phase. Then two results from the two different pipelines will be combined to make a prediction. They used F1 score in their model. F1 score considers both accuracy and recall rate, therefore it can better indicate the model's health. In statistics, the F1 score is a measure of how accurate a dichotomous (or multitasked dichotomous) model is. It takes into consideration the categorization model's accuracy and recall. With a maximum of 1 and a minimum of 0, the F1 score may be thought of as a weighted average of model accuracy and recall rate. A higher number indicates a more accurate model.

Md. Touhidul Islam1, Sanjida Reza Rafa2 Md. Golam Kibria[11] proposed a system for early prediction of heart disease using principle component analysis, hybrid genetic algorithm and k-means algorithm. Their research aims to detect heart disease early on. Their strategy can be summarized in a few phases that are implemented to improve the final clustering. They use PCA and an unsupervised heuristic k-means approach with metaheuristic Genetic Algorithms to minimize the dimensionality of the dataset for improved combinatorial optimization. Their suggested algorithm's clustering quality has substantially increased after convergence. The accuracy of predicting heart disease in its early stages was 94.06% in the end.

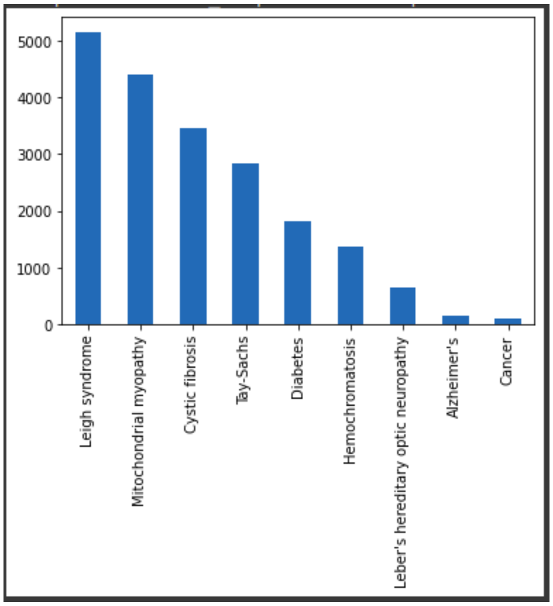
Hala Ahmed, Hassan Soliman and Mohammed Elmogy [9] proposed a system that applies multiple machine learning algorithms to identify genetic biomarkers associated with Alzheimer's Disease.They used three machine learning methods, including random forest, support vector machine and others and compared them to determine the ML classifier's precision. Based on what they have studied, Alzheimer's Disease is a complex disorder and It is critical to predict the likelihood of future disease outbreaks. For their whole genome approach random forest, and support vector machine algorithms were applied to all genetic data. The total accuracy of both respectively were 97.97 %, 95.88 % according to the data. This suggests that machine learning is an effective or promising tool for Alzheimer's Disease classification. As a result, deep learning will be required in the future for SNPs detection of disease-related biomarkers, along with early prediction and diagnosis with high classification precision.

4 What is new in the Proposed Project?

Many of us pass on mutated genes to our offspring without realising it. We don't know what the mutated genes in our bodies are because we don't know what they are. So, the proposed project's novel feature is that it will assist all people before they marry in determining the likelihood of passing on defective genes to their children in the future.

5 Proof of concept

- imports library that we will use it
- dataSet-train = reading of training dataSet using pandas library
- Display our training dataSet using
- dataSet-train.head()
- dataSet-test = reading our testing dataSet using pandas library
- Display our testing dataSet using
- dataSet-test.head()
- Drop non needed columns that will not help us
- dataSet-train.drop(columns)
- dataSet-test.drop(columns)
- Display our training and testing data after removing non needed columns
- dataSet-train.head()
- dataSet-test.head()
- Display our target graph to check if our dataSet balanced or not
- dataSet-train[target].value-counts().plot.bar()



- Make a pre processing on our training dataSet and testing dataSet using get-dummies function
- dataSet-train[column] = pandas.get-dumies(dataSet-train[column])
- dataSet-test[column] = pandas.get-dumies(dataSet-test[column])
- Prepare training and testing data
- x-train, x-test, y-train, y-test = dataSet-train.drop(target column, axis=1),
- dataSet-test, dataSet-train[target column], dataSet-train[target column]
- printing to see training and testing data

KNN

- Apply the KNN algorithm to create our model
- model = KNeighborsClassifier(Number of neighbours = 10)
- Let’s train our model
- model.fit(x-train, x-test)
- Let’s test our model
- model.predict(x-test)
- Now we will test accuracy of KNN algorithm
- model.score(x-test, model.predict(x-test))
- **Accuracy of KNN: 0.5712625462229266**

6 Project Management and Deliverables

6.1 Deliverables

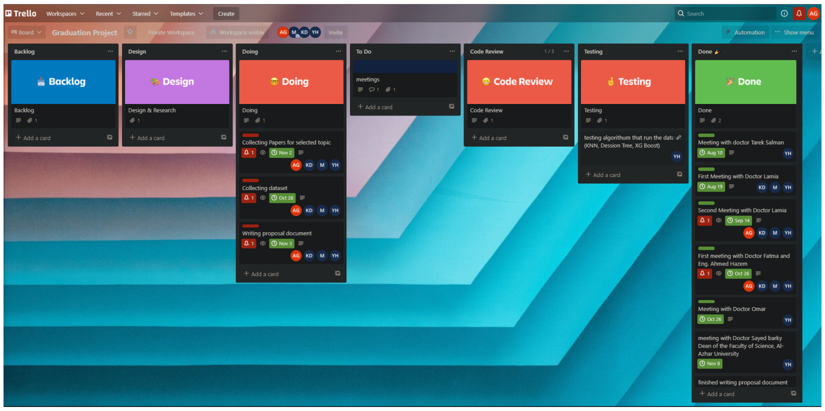
- The project will provide a model for everyone to help them understand what genetic diseases willaffect their children's future generations

- Describe in brief detail the features of each of the deliverables. Our model uses data from a single or altered gene, which is analysed by our machine learning system to determine how high a human could be infected.

- Project milestones
  - 1- Collect a useful dataset containing snippets of altered genes OR the entire genome of patients with genetic diseases.
  - 2- Make a mobile app that allows anyone to enter their entire genome or snippets of mutated genes.
  - 3- Create a robust model for predicting genetic illnesses in future generations.
  - 4- Trying out a variety of algorithms to find the best one for building the model on top of.
  - 5- Deploy our software in the marketplace.

6.2 Tasks and Time Plan

Trello (Task Plan): https://trello.com/b/JdLsASKM/graduation-project



Backlog
- Backlog
- Backlog
- + Add a card

Design
- Design
- Design & Research
- + Add a card

Doing
- Doing
- Doing
- Collecting Papers for selected topic — 1 — Nov 2 — AG KD M YH
- Collecting dataset — 1 — Oct 28 — AG KD M YH
- Writing proposal document — 1 — Nov 3 — AG KD M YH
- + Add a card

To Do
- meetings — 1 — 1
- + Add a card

Code Review 1 / 3
- Code Review
- Code Review
- + Add a card

Testing
- Testing
- Testing
- testing algorithum that run the dataset (KNN, Dession Tree, XG Boost) — YH
- + Add a card

Done
- Done
- Done — 2
- Meeting with doctor Tarek Salman — Aug 10 — YH
- First Meeting with Doctor Lamia — Aug 19 — KD M YH
- Second Meeting with Doctor Lamia — 1 — Sep 14 — AG KD M YH
- First meeting with Doctor Fatma and Eng. Ahmed Hazem — 1 — Oct 26 — AG KD M YH
- Meeting with Doctor Omar — Oct 26 — YH
- meeting with Doctor Sayed barky Dean of the Faculty of Science, Al-Azhar University — Nov 8 — YH
- finished writing proposal document
- + Add a card

7 Supportive Documents

Add sections covering one or more of the following:

- Dataset.

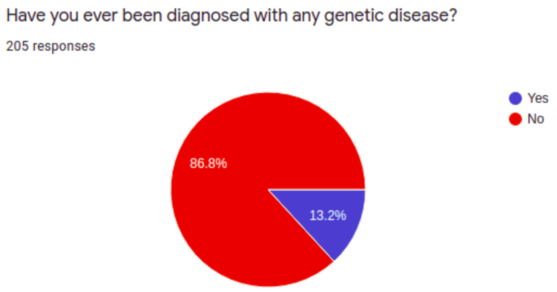
- Until now, we've been using a questionnaire data set. In addition, we apply three algorithms KNN.

- Accuracy of KNN with DataSet -> 0.5712625462229266

- Contact documents

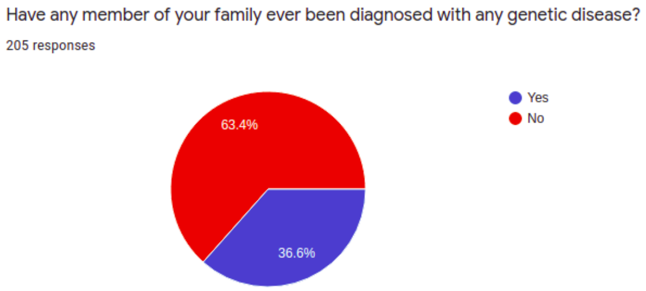
- users/survey

- we collect (205) responses in two days some people from Saudi Arabia, Dubai. Here is our statistics.



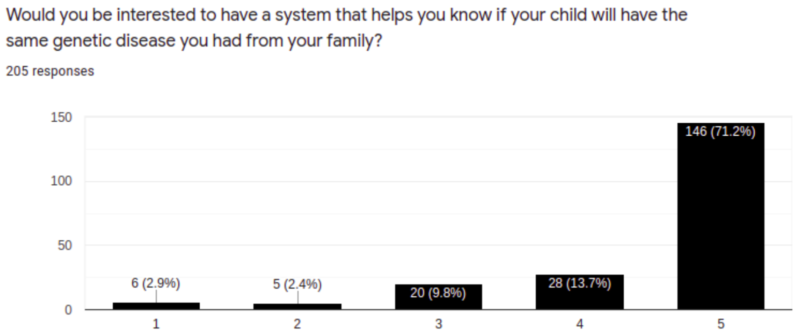
Have you ever been diagnosed with any genetic disease?

205 responses



Have any member of your family ever been diagnosed with any genetic disease?

205 responses



Would you be interested to have a system that helps you know if your child will have the same genetic disease you had from your family?

205 responses

- Contacting authors.
- we met Dr. Tarek Salman, who sent me to Dr. Ezzat Elsobky, who ran a genetic disease testing laboratory. we also had a meeting with the Dean of Al-Azhar University's Faculty of Science.

References

[1] James F Gusella et al. “A polymorphic DNA marker genetically linked to Huntington’s disease”. In: *Nature* 306.5940 (1983), pp. 234–238.

[2] M. T. Mira Kania Sabariah, S. T. Aini Hanifa, and M. T. Siti Sa’adah. “Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)”. In: *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*. 2014, pp. 238–242. DOI: 10.1109/ICAICTA.2014.7005947.

[3] Manasee Kurkure and Anuradha Thakare. “Lung cancer detection using Genetic approach”. In: *2016 International Conference on Computing Communication Control and automation (ICCUBEA)*. 2016, pp. 1–5. DOI: 10.1109/ICCUBEA.2016.7860007.

[4] Frangly Francis and T. N Namitha. “Ensemble Approach for Predicting Genetic Disease through Case-Control Study”. In: *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. 2018, pp. 326–330. DOI: 10.1109/ICICCT.2018.8473216.

[5] Joseph M. De Guia, Madhavi Devaraj, and Larry A. Vea. “Cancer Classification of Gene Expression Data using Machine Learning Models”. In: *2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology,Communication and Control, Environment and Management (HNICEM)*. 2018, pp. 1–6. DOI: 10.1109/HNICEM.2018.8666435.

[6] Umid Kumar Dey and Md. Sajjatul Islam. “Genetic Expression Analysis To Detect Type Of Leukemia Using Machine Learning”. In: *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*. 2019, pp. 1–6. DOI: 10.1109/ICASERT.2019.8934628.

[7] Wei Li et al. “Gene expression value prediction based on XGBoost algorithm”. In: *Frontiers in genetics* 10 (2019), p. 1077.

[8] Dina Y. Mikhail. “Pre-cancer Diagnosis via TP53 Gene Mutations by Using Bioinformatics amp; Neural Network”. In: *2019 International Engineering Conference (IEC)*. 2019, pp. 136–141. DOI: 10.1109/IEC47844.2019.8950565.

[9] Hala Ahmed, Hassan Soliman, and Mohammed Elmogy. “Early Detection of Alzheimer’s Disease Based on Single Nucleotide Polymorphisms (SNPs) Analysis and Machine Learning Techniques”. In: *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. 2020, pp. 1–6. DOI: 10.1109/ICDABI51230.2020.9325640.

[10] Yifan Gao et al. “An adaptive machine learning pipeline for predicting the recurrence of gastric cancer”. In: *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*. 2020, pp. 408–411. DOI: 10.1109/ISCTT51595.2020.00076.

[11] Md. Touhidul Islam, Sanjida Reza Rafa, and Md. Golam Kibria. “Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means”. In: *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. 2020, pp. 1–6. DOI: 10.1109/ICCIT51783.2020.9392655.

[12] Nibeth Mena Mamani. “Machine Learning techniques and Polygenic Risk Score application to prediction genetic diseases”. In: *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal* 9.1 (2020), pp. 5–14.

[13] J. Neelaveni and M.S.Geetha Devasana. “Alzheimer Disease Prediction using Machine Learning Algorithms”. In: *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2020, pp. 101–104. DOI: 10.1109/ICACCS48705.2020.9074248.

[14] Mayuri Mehta et al. *Tracking and Preventing Diseases with Artificial Intelligence*. 2021.

[15] Karthik Sekaran and M Sudha. “Predicting autism spectrum disorder from associative genetic markers of phenotypic groups using machine learning”. In: *Journal of Ambient Intelligence and Humanized Computing* 12 (2021), pp. 3257–3270.