Journal of Computer Science

Original Research Paper

HMGD: A High-Accuracy Model for Detection and Prediction of Respiratory Genetic Diseases

[1]Kamal ElDahshan, [2]Hesham Hefny and [2]Iman Ahmed ElSayed

[1]Department of Computer Science, El Azhar Univeristy, Cairo, Egypt
[2]Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo, Egypt

Article history
Received: 23-08-2023
Revised: 30-10-2023
Accepted: 08-11-2023

Corresponding Author:
Iman Ahmed Elsayed
Department of Computer Science, Faculty of Graduate Studies for Statistical Research, Cairo, Egypt
Email: support@thescipub.com

Abstract: Respiratory genetic diseases are considered a major participant in the reasons of death worldwide nowadays and were one of the major participants in helping in increasing the numbers of COVID-19 patients. It is considered one of the most alarming diseases affecting in particular the respiratory system. The journey of early detection of respiratory genetic diseases is considered to be very challenging today to assist in lessening the percentage rate of death since people with these diseases are more vulnerable to being infected by COVID-19 and other dangerous diseases than others. Also, it is considered a very difficult mission for medical practitioners because of the high requirement for expertise and knowledgeable practitioners. While, predicting or detecting respiratory genetic disease in an early phase has many gaps and lacks accuracy accommodated with speed as well; as a result any slight update in the accuracy accommodating speed will be considered of great improvement and importance which will later result in the reduction of the increasing number of genetically diseased patients as the well-known diseases of Alpha-1 antitrypsin deficiency, Cystic fibrosis, Kartagener syndrome and many other respiratory genetic diseases. In this study we will introduce a new hybrid-model approach (HMGD) based on merging two outstanding soft computing optimization algorithms which weren't used before in neither detection nor prediction of diseases which are Extended Compact Genetic Algorithm (ECGA) and Compact Co-Evolutionary Algorithm (CCoEA); one for which ECGA will act for the feature selection phase and output will be fed to the CCoEA for feature optimization resulting in the certainty factor of the detected/predicted respiratory genetic disease. The model will be used through a graphical user-friendly interface built up especially for the model to analyze data, learn from that output data, and result in a tactile and touchable prediction/detection for the respiratory genetic disease. The HMGD model proved its reliability and outstanding performance over other known computational models by an accuracy of 98.27% for respiratory genetic diseases' prediction in 1.03 sec, while an accuracy of 97.89% for respiratory genetic diseases' detection in 1.4 sec. The model proved to achieve a higher level of accuracy in the detection or prediction of respiratory genetic diseases than other machine learning models.

Keywords: Respiratory Genetic Diseases, COVID-19, Evolutionary Computation, Genetic Algorithm, Co-Evolutionary Algorithms, Machine Learning Models, Disease Prediction, Detection, COPD Disease, Protein Sequence

Introduction

COVID-19 is continuing to spread rapidly worldwide, generating a wake of devastating health, economic, and social consequences, and remains the leading cause of death globally. This urgency of the situation has simultaneously driven eyes to those patients mainly having respiratory genetic diseases as they are more

vulnerable to COVID-19. If we consider for example one of the most significant respiratory genetic diseases which is Alpha-1 Antitrypsin Deficiency (A1AD or AATD); as to illustrate and emphasize the need for a high accuracy detection/prediction computational model. Accompanying Down syndrome and cystic fibrosis, AATD is one of the three most commonly recognized, potentially fatal, and scarce diseases affecting those of European descent (Fleming, 2021). The frequency of its occurrence is 1 in every 1500-3,500 individuals with European ancestry. As for the United States of America, AATD affects 1 in every 3000-5000 people. That is approximately 70,000-100,000 individuals are estimated and considered to be critically affected by AATD in the US. With an approximate number of 175 million cases internationally, COPD holds a rank of 38th on the World Health Organization's global prevalence for disease and 4th place in the death charts in Fig. 1 and if we take a deeper look, AATD contributes to 1% of these COPD cases also. However, AATD is underdiagnosed or misdiagnosed. It is projected that less than 10% of these individuals have obtained an accurate diagnosis. Many individuals showing symptoms of AATD are diagnosed with asthma or Chronic Obstructive Pulmonary Disease (COPD), but unfortunately, they don't undergo AATD tests with long delays (mean of 5-8 years) to reach an exact diagnosis if it is made at all. In addition to all of the previous, COVID-19 vaccination is not effective for these patients and as a result, these patients are 8.8% more infected by COVID-19.

For diagnosis of such diseases, investigations such as chest X-rays and spirometry are generally performed, however more complex procedures such as Bronchoscopy, Plethysmography, Computerized Tomography (CT) or Positron Emission Tomography (PET) scans may also be performed. These investigations require lot of time, effort and money to be analyzed. Moreover, supervising and regulating medicines gets delayed which sometimes may have unpleasant effect on the patient.

Based on what is mentioned above, new computational methods such as optimization algorithms and evolutionary computation have appeared to help in reducing the timing required for such tests, results will be more accurate and it will be much easier untangle, find and resolve most complicated cases.

The objective of our work is to present a new hybrid approach for approaching the early detection of respiratory genetic disorders and anomalies from protein sequences (our case study) with better accuracy than other used models and in less time as time is very critical in this case study. This approach involves merging both the Extended Compact Genetic Algorithm (ECGA); a learning optimization evolutionary algorithm; and the Compact Co-evolutionary algorithm (CCoEA); an optimization Co-evolutionary algorithm. The dataset used was collected from the universal protein knowledge bank.

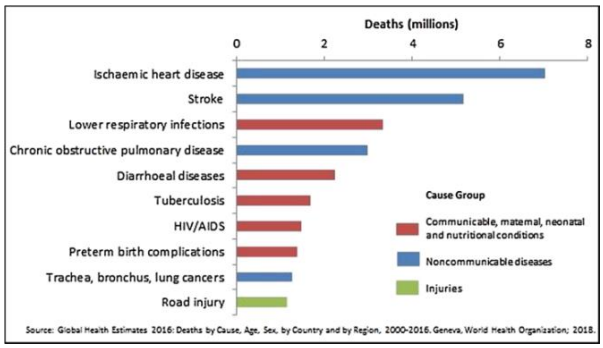

Fig. 1: Respiratory diseases contributing in deaths worldwide

The hybrid model starts the feature selection process for the given protein sequence (either has an anomaly or not) using ECGA with standardized parameters that best fit our problem (detection/prediction), while CCoEA is responsible for feature optimization with also certain standardized parameters that fit most finding the certainty factor for the predicted/detected respiratory genetic disease.

The following manuscript's work mainly has the following contribution:

1. Data is collected and preprocessed from the Universal Data Knowledge Bank; only that which we need for the case study and a database has been built up especially for it to feed the model with it
2. The usage of ECGA for feature selection for the first time with standardized parameters for a given protein sequence for the possible liability of the detection or prediction of the protein to have a respiratory genetic disorder or not
3. The results and data generated as output from the ECGA is the fed for feature optimization process using CCoEA to state and confirm if this protein sequence has a respiratory genetic disorder already and liable to it; the question at this point will be: What will be the certainty factor for this disease that has been detected and most likely to occur for the patient?
4. Combining and merging both ECGA and CCoEA acting as a one combined hybrid model; which hasn't been done for these two models before; to compete and compare with different machine learning models

Many researchers have worked in disease prediction or disease detection but almost none worked on a model that performs both disease prediction and detection. Moreover, respiratory genetic diseases remain untouched with many of disease prediction or detection computational models and techniques. As a result, e will organize our discussion of related work into two parts, firstly, anomaly-driven prediction and secondly, anomaly-driven detection methods for different diseases.

We will start first with related work to prediction of respiratory diseases: Lung cancer is one of the most

causing to death diseases in the world today and especially in developing countries. Lung Cancer is mostly caused due to some genetic factors. Various works have been done in predicting lung cancer but unfortunately they were lacking intelligent and precise prediction. Priya and Meyyappan (2021) has proposed the Lung cancer prediction system based on data mining. This study analyzed the lung cancer prediction using classification algorithm such as Naive Bayes, SVM and Random forest algorithm and the main aim of the paper was to provide the earlier warning to the users (Priya and Meyyappan**,** 2021). Perna and Tagarelli (2019) used an advanced deep-neural-network architectures approach in predicting respiratory diseases, at either level of abnormal sounds or pathology classes (Perna and Tagarelli, 2019). Koppad *et al.* (2019) a proposed in 2016 a decision Tree technique for better performance in COPD diagnosis in individual patient. The experimental results showed a promising accuracy in diagnosing COPD patient and efficiency of the proposed system Application of big data analytics in hea1thcare system to predict, 2016. In 2011, a simple decision tree method that uses variables commonly gathered by physicians provided a quick prediction of the severity of the disease in patients with stable Chronic Obstructive Pulmonary Disease (COPD) by (Esteban *et al*., 2011).

Secondly, we will illustrate the related work that was concerned with the anomaly detection of respiratory genetic diseases which are as follow: Jayaraj and Sathiamoorthy (2020) presented a new deep learning classification model for lung cancer detection. The presented model involves four main steps namely preprocessing, feature extraction, segmentation and classification. A Particle Swarm Optimization (PSO) algorithm is used for segmentation and Deep Neural Network (DNN) is applied for classification. The presented PSO-DNN model is tested against a set of sample lung images and the results verified the goodness of the projected model on all the applied images (Jayaraj and Sathiamoorthy, 2020). Spathis and Vlamos (2019), Panayiotis constructed a study that have examined the clinical decision support systems (machine learning technique) in healthcare, in particular about the detection of respiratory diseases, such as chronic obstructive pulmonary disease (COPD and its results was promising to a certain limit (Diagnosing asthma and chronic obstructive pulmonary disease with machine learning, 2019). Amaral *et al.* (2012) developed a clinical decision support system based on Machine Learning (ML) algorithms to help the diagnostic of Chronic Obstructive Pulmonary Disease (COPD) using Forced Oscillation (FO) measurements. The results of this study indicated that it may contribute to easy the diagnostic of COPD by using forced oscillation measurements (Amaral *et al*., 2012).

All of this previous related work done had a major deficiency which is that different models used were either for detecting an anomaly only or predicting the required anomaly only not combining both properties which might be needed in certain critical cases for diagnosis of respiratory genetic diseases' patients which are most vulnerable through the crisis of COVID-19.

As a result, our work was designed to fill up the gaps that are missing and propose a new hybrid model between the two algorithms (the Extended Compact Genetic Algorithm (ECGA) and the Compact Co-Evolutionary Algorithm (CCoEA)) which will be able to detect and predict respiratory genetic diseases given a certain sequence as an input to be able to handle those patients well through COVID-19. Also, a feature have been accompanied to this model that wasn't available in other models which is stating the certainty factor for each predicted disease and as a result can take in return proper medical actions in a hurry if needed.

Materials and Methods

The proposed model aims detecting/predicting respiratory genetic diseases (i.e.: Anomalies in a given sequence) at an early stage which can enable to overcome and treat them appropriately using a computational hybrid model that can assist and help medical practitioners. The proposed Hybrid Model (HMGD) is composed of four phases for the respiratory genetic disease prediction/detection process which are: Data generation and preprocessing firstly, coming next the gene sequence is given as a data entry through a Graphical User Interface (GUI) built, afterwards feature selection phase using the Extended Compact Genetic Algorithm (ECGA) and at last results from the middle phase are fed to the compact co-evolutionary algorithm for feature optimization.

Dataset Description

Initially the data set generated was composed from the universal protein data bank (UKB). The Universal Protein Source (UniProt) is an all-inclusive source for data concerned with protein sequences and annotation data.

The UniProt Knowledgebase (UniProtKB) is considered a crucial hub for different and variant functional information on proteins, with accurate, consistent and rich annotation (Breuza *et al*., 2016). Moreover, it includes widely accepted biological ontologies, databases, classifications and cross-references connected to each other. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc) in Fig. 2.

The database generated corresponds to all human genetic diseases that can be put from a practitioner to be tested further but our main interest in our case study is respiratory genetic diseases.

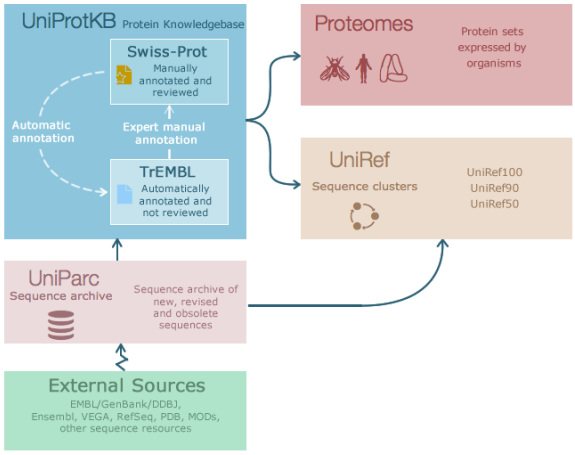

Fig. 2: UniProtein databases (UniProt, 2021)

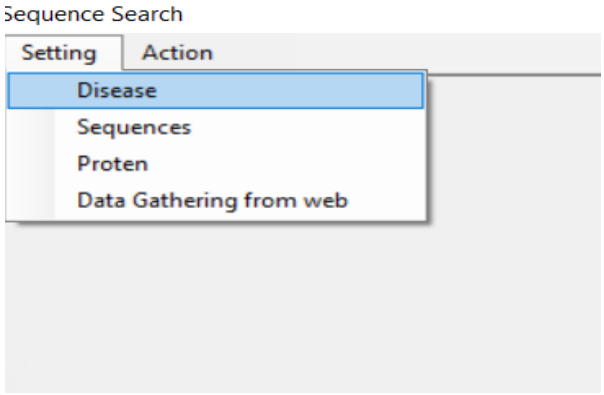

Fig. 3: data extracted from the UKB on the GUI

As it was very hard and unreachable to get real data from hospitals (unless the data has been used in a project), the data of sequences in the UniProtKB (UKB) has been used as a reference for the HMGD model since it is reliable data.

First of all data have been filtered in the UniProKB to select only those protein sequences of human not all species which resulted in 204,961 record. The dataset then has been filtered and preprocessed for checking for NULL values, duplicate values, missing values, and outliers.

A Database has been built up from the dataset extracted from the UniProKB and embedded and connected with the interface for the model.

Figure 3, an image for data extracted from the UKB on the GUI of the implementation; we either extract all human diseases for a database by its own, protein sequences for humans as a standalone database, proteins discovered and verified if needed later in any work in a database and finally all data needed can be imported as a one big database to be used for everything. The database generated can be updated on real time basis since the app is synchronized to the UKB website directly.

By theses four databases, we have covered every aspect that will either be needed for prediction or detection or both of them to be compared with the input sequence of the patient.

The database generated corresponds to all human genetic diseases that can be put from a medical practitioner to be tested further but our main interest in our case study is respiratory genetic diseases.

The Proposed HMGD Model

The proposed computational hybrid model for predicting respiratory genetic diseases through a given sequence (HMGD) that can help and assist medical practitioners; firstly in effectively identifying those respiratory genetic diseases in an early stage through the corona virus epidemic and then describing the certainty factor by which the found respiratory genetic diseases can affect a certain human being which later will help them in either determining if this patient will be in the red zone patients if he was affected by a dangerous diseases as COVID-19 or not.

Also, the proposed model will be able to detect respiratory genetic diseases thoroughly and accurately so as help practitioners in obtaining an alternative opinion and how much it's reliable in critical situations if needed through checking a patient through the epidemic of corona virus.

How the HMGD Model Functions?

Let D = {($name_1$, seq_1), ($name_2$, seq_2), ….. ($name_i$, seq_i)}be the dataset available for the model where the ordered pair ($name_i$, seq_i) presents the name of all possible genetic disease names and their corresponding sequences.

Feature Selection of HMGD

This is a step in which unneeded and unresolved data from the built up database have been eliminated and reduced in size in order to be specific and precise in our output results. This model has used the Extended Compact Genetic Algorithm (ECGA) to select which disease will be compatible with the given input sequence according to the below given equations. These equations represent two things which need further explanation. ECGA has two magnificent properties which are called "Marginal Product Models (MPMs)" and the other one is "linkage learning" which assist in solving and managing hard problems.

Firstly, the property of linkage learning assists in conserving building blocks of good properties through crossover of generations which helps later effectively in dividing the problem into small sub problems and gives a definite guarantee that the perfect result will still evolve.

Concerning the other property MPMs, it is considered a unique type of class of probabilities that assists in

gathering information which is both important and superior concerning the population. Afterwards, it samples it in the next step with a new better population. A MPM is split into two divisions; the first division is a division over the variables and the second division is a probability distribution over each of the partitions. In ECGA, MPMs is capable of representing the probability distribution for more than one gene at a time.

Moreover, ECGA embraces the concept of the Minimum Description Length (MDL) to be an effective criterion to decide how good the learned joint probability distribution is.

Phase 1: ECGA Part for HMGD Model

1. Initialize a population of size (N) randomly
2. For first generation, each (i) in (N)

- Find the fitness value of each individual
- (chromosomes)
- Perform a tournament selection of size (s)
- Build a probabilistic model for the population
- using a greedy MPM search

(The identification of MPM using MDL and the other is the creation of a new population based on MPM).

- Sample the probabilistic model generated for appearance of new individuals

The identification of MPM in every generation is formulated as a constrained optimization problem as follow:

$$Minimize: C_m + C_p \tag{1}$$

$$Subject\ to: 2^{l_{bb,i}} \le N_p \forall i \in [1, N_{bb}] \tag{2}$$

C_m model complexity and this represents the cost of a complex model given by:

$$C_m = \log_2(N_p + 1)\sum_{i=1}^{N_{bb}}(2^{l_{bb,i}} - 1) \tag{3}$$

C_p compressed population complexity which represents the cost of using a simple model as against a complex one and is evaluated as:

$$C_p = \sum_{i=1}^{N_{bb}}\sum_{j=1}^{2^{l_{bb,i}}} N_{ij}\log_2\left(\frac{N_p}{N_{ij}}\right) \tag{4}$$

3. If MPM model has converged, then terminate
4. If MPM model hasn't converged, repeat all steps on step 2 until convergence of MPM model has been obtained

In phase 1, ECGA uses the following parameters which are considered adjusted settings in order to achieve the highest accuracy of prediction or detection and selection of required diseases on all tested cases.

- Population size: 80
- Chromosome length: 40
- Tournament size: 16
- Seed: 0.254534
- Cross over probability: 1

Feature Optimization of HMGD

The proposed model will be using the Compact Co-Evolutionary Algorithm (CCoEA). This novel algorithm was introduced mainly combined from both Compact Evolutionary Algorithm (CEA) and the Co-Evolutionary Algorithm (CoEA) for the purpose of improving convergence by overcoming bias impprovement of solutions and reducing both runtime and memory consumption.

CCoEA works with two vectors; one called the better probability vector (PV_{better}) and the other one is called the worst probaility vector (PV_{worst}). The PV_{better} is a probability vector mainly focused on exploitation process in order to be dedicated just for increasing the speed of convergence to the better solution, while PV_{worst} is more focusing and aiming at preventing the premature convergence. Thus, exploitation strategy is applied on PV_{better} while exploration strategy is applied on PV_{worst}. If PV_{worst} finds a more effective solution in its probability vector rather than that in PV_{better}, it is directly switched with PV_{better}. By that, these two PVs keep cooperating all the way till the evolving process is ended resulting in preventing the premature convergence and improve the convergence speed to the required result as well.

Exploitation and Exploration Strategies in CCoEA in HMGD model

Exploitation mainly focuses on generating different individuals to search the vast range of the X_{elite}.

On the other hand, exploration preserves and keeps up the variety of the population throughout the probabilities in the PVs. Since the boundaries of our probability is between [0,1], thus when the elements of PV_{worse} is near to 0 or 1, then PV cannot be improved as the individuals generated will be almost the same. But, when the elements of PV is close to 0.5, the individuals generated will be different and various.

Therefore; in our HMGD model, PV_{worse} will be defined to 0.5 only if:

(1) When the values of PV_{worse} are either larger than 0.8 or smaller than 0.2
(2) When the $X_{elite.}$ generated by PV_{worse} cannot be improved for M generations
(3) When the difference between PV_{worse} and PV_{better} are smaller than α, where α is to be as follow:

$$\alpha = \frac{\sum_{i=1}^{L}(PV_{better}^{i} - PV_{worse}^{i})}{L}$$

L = Number of *PV* elements
$PV_{better}^{i} - PV_{worse}^{i}$ = Refer to the element I in their
= *PV* respectively
m = 10 for obtaining the best results
α = 0.2 for obtaining the best results

CCoEA is considered to be a cooperative co-evolutionary algorithm since two PVs cooperate with each other in the whole evolving process of CCEA to trade off the exploitation and exploration, which can prevent the premature convergence and improve convergence speed.

This part of the HMGD model will help in optimizing and improving the resultant from the ECGA that is given as input to the CCoEA to tune it and find the certainty factor for each respiratory genetic disease that matches most the sequence of the patient and this will help a lot in the accuracy of the respiratory diseased patient's treatment through Covid-19 or any other crisis.

Phase 2: CCoEA part for HMGD Model

Step (1) = Initialization:
Step (1.1) = Initialize generation = 1
Step (1.2) = Initialize PV_{better} and PV_{worse} by setting all the probabilities inside as 0.5
Step (1.3) = Generate two individuals $elite_{better}$ and $elite_{worse}$ through PV_{better} and PV_{worse}, respectively

After initialization, the algorithm process in phase 2 will divided into 3 steps: Step 2.1(exploitation), step 2.2 (exploration) and step 2.3 (competition).

Step 2) Update PVs:
Step 2.1) Exploitation:

- Calculate the Virtual Population VP
- Apply the exploitation strategy on $elite_{better}$ to obtain $indiv_{localBest}$
- [winner,loser]=compete ($elite_{better}$,$indiv_{local\ Best}$).
- if (winner == $indiv_{local\ Best}$)
  $elite_{better}$ = $indiv_{local\ Best}$

End if

- for(i = 0; i < num; i++)
  If (winner[i]==1)
  $PV_{better}[i] = PV_{better}[i] + V\ P$
  Else
  $PV_{better}[i] = PV_{better}[i] - V\ P$
  End if

End for
Step 2.2) Exploration:

- Applying the exploration strategy to update PV_{worse}.
- Generate $indiv_{new}$ by means of PV_{worse}.
- [winner,loser]=compete($elite_{worse}$,$indiv_{new}$).
- if (winner == $indiv_{new}$)
  $elite_{worse}$ = $indiv_{new}$

End if

- For (i = 0; i < num; i++)
  If (winner[i]==1)
  $PV_{worse}[i] = PV_{worse}[i] + V\ P$
  Else
  $PV[i]_{worse} = PV_{worse}[i] - V\ P$
  End if

End for

Step 2.3) Competition between PV_{better} and PV_{worse}
[winner, loser] = compete ($elite_{better}$, $elite_{worse}$)
If (winner == $elite_{worse}$)
switch PV_{better} and PV_{worse}
switch $elite_{better}$ and $elite_{worse}$
End if

Step 3) Termination:
If the termination condition is met which is;
!! !! ((max generation is reached) or ($elite_{better}$) == 1 or $elite_{better}$ is equal to $elite_{worse}$ and at the same time, all elements in PV_{better} and PV_{worse} are either 1 or 0)
stop and output $elite_{better}$
Else
generation = generation + 1

go to Step 2) (continue evolving of generation process)
End if

In phase 2, CCoEA uses the following parameters which are considered adjusted settings in order to achieve the highest accuracy for the certainty factor of the diseases on all tested cases.

- Minimum virtual population: 0.02 (2% of the given population in PV_{better})

NB: If the value of the min.virtual population is too big, the algorithm in this phase will tend to exploit sooner

than it must and premature convergence will occur. Also, if it is too small, exploration strategy will take too much time and convergence might not occur at all).

Maximum virtual population: 0.2 (20% of the given population in PV_{better}).

NB: if the value of the max. virtual population is too big, the algorithm in this phase will tend to converge sooner than it must before the virual population does. Also, if it is too small, convergence might not occur at all.

- $\alpha = 0.2$
- M = 10
- Maximum number of generations = 300

Results

The below section provides a comparative inspection study of the results that were attained by the suggested approach. A hybrid model for predicting genetic diseases from a given sequence has been developed. It is mainly based on the combination of two soft computing algorithms which are ECGA and CCoEA. Mentioned below are the tables providing a comparative study of the results that were attained by the suggested approach with respect to accuracy and execution time for both prediction and detection.

The results were built with hardware configurations to run the hybrid algorithm as follow:

- Processor: Intel core i7 VPRO- 4810MQ CPU
- CPU Speed 2.80 GHz
- RAM Capacity: 24 G

Tables 1-2 values show precisely that the presented HMGD model reveals a highly ranked performance to reach the $elite_{better}$ result by acquiring the maximum accuracy of 98.27% in prediction and 97.89% accuracy through detection.

Table 1: Results for proposed model in prediction compared to other used models before

Model for prediction	Accuracy %	Execution time in min
Proposed (HMGD)	98.27	1.03
Naïve Bayes	89.90	2.60
Recurrent neural network	74.00	1.89
Decision tree	74.03	2.05

Table 2: Results for proposed model in detection compared to other used models before

Model for detection	Accuracy %	Execution time
Proposed (HMGD)	97.89	1.40
Random forest	97.70	2.03
Artificial neural network	95.00	2.33
Decision tree	80.30	1.97

Discussion

Simultaneous to the HMGD model, the Naïve Bayes model tried to handle the prediction process by obtaining a moderate accuracy value of 89.90% and execution time of 2.6 sec. On the other hand, the decision tree model shows less performance by acquiring the accuracy value of 74.03% and an execution time of 1.89 sec. However, recurrent neural network technique does not show its effective performance. It showed less efficiency by achieving minimum accuracy values of 74%.

Figures. 4-8 and tables, it is clear that the best results are obtained by the presented HMGD model either in prediction or detection. It is completely verified from the maximum accuracy and execution time is for our model can be applied as an automated diagnosis and prediction tool for respiratory genetic diseases.

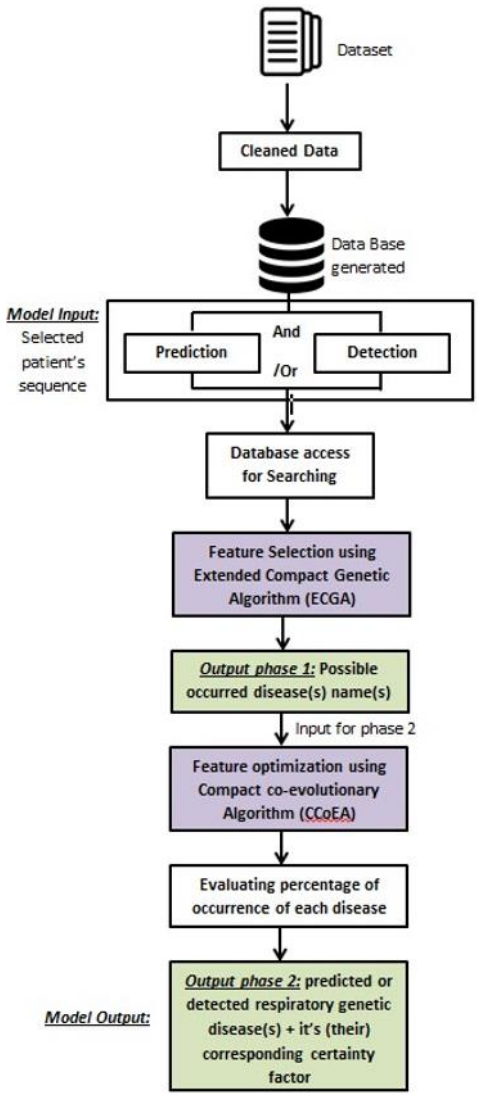

Fig. 4: The HMGD model technique

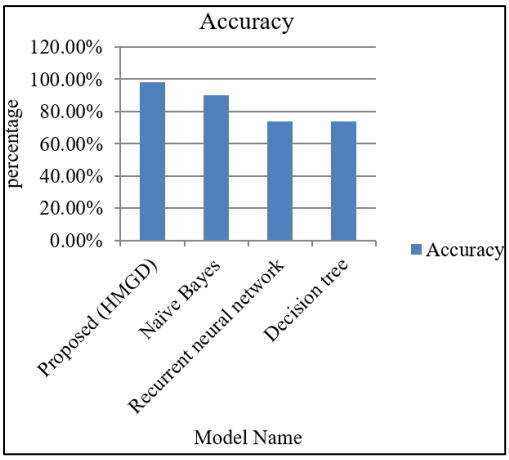

Fig. 5: Comparative analysis of the accuracy of proposed model through prediction compared to other existing models in literature review

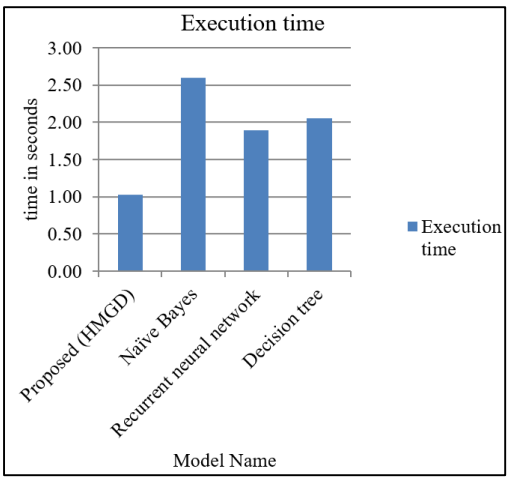

Fig. 6: Comparative analysis of the execution time of proposed model through prediction compared to other existing models in literature review

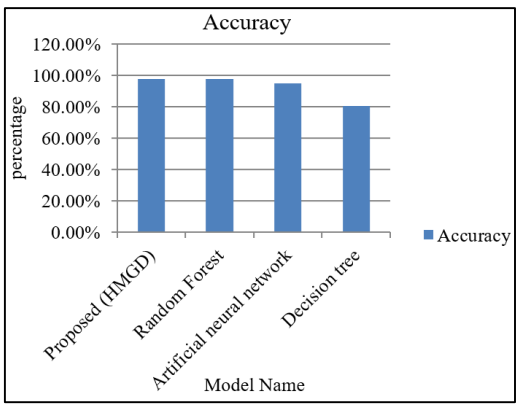

Fig. 7: Comparative analysis of the accuracy of proposed model through detection compared to other existing models in literature review

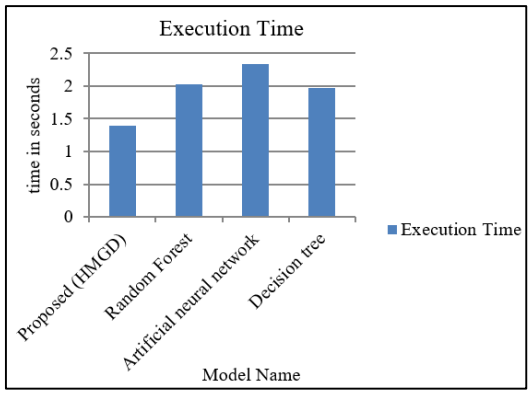

Fig. 8: Comparative analysis of the execution time of proposed model through detection compared to other existing models in literature review

Conclusion

In this manuscript, a Hybrid Model (HMGD) for respiratory genetic diseases prediction and detection using a combination of two evolutionary algorithms which are ECGA and CCoEA has been proposed and discussed. The proposed model HMGD is extremely precise and authentic and is able to predict and detect the respiratory genetic disease quickly.

The proposed model is a combination of feature selection and feature optimization. Feature selection is used mainly in the model to eliminate non-needed data, increases accuracy and improves understanding of results by obtaining possible disease names matching the sequence given as an input to pass it later to the next phase which is feature optimization. Feature Optimization is applied to fine the certainty factor for each matched disease which will make the results more optimized. For this purpose, ECGA and CCoEA are used. ECGA matches the given input sequence with the database and finds the most equivalent respiratory genetic diseases while CCoEA being a computational method optimizes the solution from the phase before it and mentions the certainty factor for each match which later provides a most candidate solution for the disease with regard to a given input sequence. This unique merge resulted into constructing a strengthen model imposing on the existing models.

The proposed model may assist the medical practitioners in early detection and prediction of respiratory genetic diseases through the COVID-19 pandemic to help those patients in particular as until now the COVID-19 various vaccines doesn't have any effect on them and doesn't give them any protection from COVID-19.

Finally, the newly proposed model; HMGD is a pioneer in composing both ECGA and CCoEA together to form a

hybrid model that output results in its prediction phase as a set of order pairs {(disease, CF for the disease), ….}.

Limitations and Future Work

In further research, other algorithms may be used other evolutionary algorithms may be used (such as Ant colony and Shuffled frog leaping algorithms, rat algorithm) over the model so as to compare the results with the results from this study and find out faster and better optimization for this model if it can exist and a whole comparative study can be performed to find better optimization techniques through evolutionary algorithms.

Also, clinical trials can be done for different genetic diseases and compared to results from our proposed model as more detailed study.

Moreover, the model can be extended to include all genetic, pandemic and non-genetic diseases that will help through the pandemic diseases' spread such as Black fungus disease and many other diseases.

Acknowledgment

Thank you to my professors and teachers who are also participant authors in this research article prof. Kamal el Dahshan and Prof. Hesham Hefny for always guiding and helping me throughout this article and my PhD thesis work. Thank you to the publisher for their support and help in the publication of this research article. We are very grateful for the resources and comments provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

The authors haven't received and financial funds to report for this article.

Author's Contributions

Kamal ElDahshan and Hesham Hefny: Mentoring and revised.

Iman Ahmed ElSayed: Implemented and tested the model and output results.

Ethics

This article is original and genuine and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and no ethical issues are involved.

References

Amaral, J. L., Lopes, A. J., Jansen, J. M., Faria, A. C., & Melo, P. L. (2012). Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer Methods and Programs in Biomedicine*, *105*(3), 183-193. https://doi.org/10.1016/j.cmpb.2011.09.009

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M. L., Magrane, M., Tognolli, M., ... & UniProt Consortium. (2016). The UniProtKB guide to the human proteome. *Database*, *2016*, bav120. https://doi.org/10.1093/database/bav120

Esteban, C., Arostegui, I., Moraza, J., Aburto, M., Quintana, J. M., Pérez-Izquierdo, J., ... & Capelastegui, A. (2011). Development of a decision tree to assess the severity and prognosis of stable COPD. *European Respiratory Journal*, *38*(6), 1294-1300. https://doi.org/10.1183/09031936.00189010

Fleming, S. (2021). These are the top 10 global causes of death - but two diseases are in decline. *World Economic Forum Global Health.* https://www.weforum.org/agenda/2021/02/top-10-global-causes-death/.

Jayaraj, D., & Sathiamoorthy, S. (2020). Deep neural network based classifier model for lung cancer diagnosis and prediction system in healthcare informatics. *In Intelligent Data Communication Technologies and Internet of Things: ICICI 2019*, (pp. 492-499). Springer International Publishing. https://doi.org/10.1007/978-3-030-34080-3_56

Koppad, S. H., Kumar, S. A., & Rao, M. K. N. (2019). *Contemplation of Computational Methods and Techniques to Predict COPD*, pp. 538-545. https://doi.org/10.1007/978-3-030-34080-3_61

Perna, D., & Tagarelli, A. (2019, June). Deep auscultation: Predicting respiratory anomalies and diseases via recurrent neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems* (*CBMf* Contemplation of Computational Methods an*S*), (pp. 50-55). IEEE. https://doi.org/10.1109/CBMS.2019.00020

Priya, T., & Meyyappan, T. (2021). Disease prediction by machine learning over big data lung cancer. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 16-24. https://doi.org/10.32628/CSEIT206669

Spathis, D., & Vlamos, P. (2019). Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics Journal*, *25*(3), 811-827. https://doi.org/10.1177/1460458217723169

UniProt, C. (2021). UniProt. *UniProt-KB.* https://www.uniprot.org/help/uniprotkb