Chronic Disease Prediction: AI Based platform

Devang Vasani
Master of Electrical and Computer Engineering
Carleton University
Ottawa, Canada
devanghasmukhbhaivas@cmail.carleton.ca

Abstract: The paper presents a health programme that aims to avoid chronic illnesses and maintain people's health. It tracks people's movements and their environment in real time by utilising wearables and sensors. After two years of testing on 1,667 people, the system was able to correctly identify 386 aberrant health occurrences. Models that accurately predicted illnesses including obesity, panic disorder, and lung disease were trained using sophisticated computer algorithms. The study demonstrates the strong correlation between environmental and lifestyle variables and health outcomes, underscoring the need of collecting extensive data to make more accurate forecasts. Additionally, the study offers a useful, low-cost model for forecasting health risks that simply makes use of a small number of characteristics.

Keywords – DNN, Feature Engineering, API, the shap module, AECOPD, COPD, Hyperparameter, ReLU, batch normalization, gradient

I. INTRODUCTION

This paper is totally based on how environments, and outer conditions effect on patients' chronical disease after patients discharged from hospital. When patients are in the hospital, they heal rapidly with the assistance of doctors and advanced medical care; nevertheless, once they are discharged, problems arise due to their surroundings and way of life. Here The three main chronic illnesses that we addressed namely obesity, panic disorder, and chronic obstructive pulmonary disease (COPD), which can occasionally be the main cause of mortality and disability. The purpose of the study is to develop an artificial intelligence (AI) platform that will allow physicians to remotely monitor patients who suffer from chronic illnesses, as well as to collect lifestyle and environmental aspect data from several sources and forecast the likelihood of developing those conditions beforehand.

II. GENERAL FLOW OF RESEARCH::

In this paper we need to make a machine learning model that predicts chronical diseases and whole this process consists of 3 steps:

1. User interface
    a. NTU medical genie smartphone app (for user interface and see results, data and communication to users and doctors)
    b. NTU medical genie platform
2. Data collection
    a. Wearable device (for hardware compatibility and data collection)
    b. Air quality sensor device (For checking environmental factors affecting health)
    c. Open environmental data API
3. Model building
    a. Pulmonary disease prediction model
    b. Panic attack prediction model
    c. Obesity prediction model

III. DATA COLLECTION

In this research data is collected through two different resources 1. Wearable device, 2. Air quality sensor device.

1) **Wearable device**:
Through Fitbit and other devices, patients live data such as physical activities, heart 211 rate, SpO2, and sleep patterns has been collected and automatically stored to the health database which in future gone used for model building.
2) **Air Quality sensing device**:
Environmental condition is also a factor that affects chronic disease such as asthma). In this study Edimax 225 Airbox has been used to collect those environmental factors (such as temperature, humidity). Those data are automatically transferred to health dataset through wireless network every 15 minutes.
3) **Open environment data API**:
In this step, to understand a patient's environmental risk, data is collected from the nearest environmental monitoring station. This step provides the groundwork for any data analysis later in the project or research. After the information is gathered, it may be analyzed to find trends, correlations, or insights that could help determine how the environment affects patients' health.

IV. USER INTERFACE

For total transparency, a user interface is necessary so that patients may access their information and consult with physicians. With the use of artificial intelligence, the NTU Medical Genie Platform (which is telecare system) collects large amounts of patient data using a range of technical methods. Effective communication between physicians and patients is made easier by the use of trend charts and pertinent data. moreover, Real-time warning features are integrated into the platform to notify medical staff when necessary. These capabilities are based on preset thresholds for abnormal vital signs that initiate health risk computation. Furthermore, it determines a person's health risks using collected data and modular illness prediction models (which researchers will develop in this next study), aiding in precision health management and the avoidance of chronic diseases.

V. METHODOLOGY

In this stage, as we have already gathered all the data by using the above approaches, the study has entered the model-building phase. Here, researchers created three different machine learning models to predict three distinct diseases: obesity, panic disorder, and COPD (one for pulmonary disease). This is because various diseases may be predicted using different characteristics or different types of variable settings in ML.

1) **pulmonary disease prediction model:**

   This disease highly correlated with people’s lifestyle and environmental data so, data related to those two variables imported before going into model building phase. During the study, they recruited 177 patients diagnosed with COPD.

   For this model, the author implements 5 machine learning model to check which one gives them best result.

   - Here random forest was implemented with 300 estimator (300 decision tree), each tree has minimum 4 split and maximum depth of tree is 30.
   - Another algorithm they implemented is a decision tree with a minimum of 1 leaf node and minimum 2 split at child node.
   - They implement linear discriminant with Lsqr solver and auto shrinkage.
   - AdaBoost was implemented using 45 different estimator and set learning rate to 1.
   - Using a 5-layer neural network with three hidden layers (each containing 45 neurons), one input layer (consisting of 27 features), and one output layer with a sigmoid activation function, the fifth model is a deep neural network that produces a probabilistic output with a range of 0 to 1. ReLU activation function and batch normalization is used in hidden nodes to normalize the inputs of the layers by re-cantering and re-scaling, which speeds up and stabilizes artificial neural network training[2].For optimization , In backward propagation, weights have also been updated using the Adam optimizer. They multiplied the adaptive learning rate by 0.1 per 60 epochs for this model. Additionally, they use the class_weights methods from the keras library to classify data points and ensure that even minority datapoint take into account during training.

2) **Panic attack prediction model**

   The goal of building this model is to build a model which predicts chances of panic attack in within next 7 days. In this paper six different algorithms proposed to check which one is good.

   - random forest was implemented with 100 estimator (100 decision tree), each tree has minimum 2 split and maximum depth of tree is 1.
   - Another algorithm they implemented is a decision tree with minimum 1 leaf node and minimum 2 split at child node.
   - They implement linear discriminant with Lsqr solver and auto shrinkage.
   - They implement regularized discriminant analysis with 1000 maximum leaf node, test interval is 100.
   - AdaBoost was implemented using 50 different estimator and set learning rate to 1.
   - The fifth model, a deep neural network, is a 6-layer neural network that has four hidden layers: one input layer (which has 61 features and is made up of continuously measured lifestyle data collected via wearable devices and environmental data centers), one output layer (which has sigmoid activation function to generate a probabilistic output ranging from 0 to 1), and four hidden layers which has 128 neurons with ReLU activation function ,256 neurons with ReLU activation function, 128 neurons with ReLU activation function, and 64 neurons with ReLU activation function respectively . This model also uses batch normalization to improve training speed and stability. Additionally, they employed the Adam solver for optimization, and for calculating loos function BCA (binary cross entropy) has been used. The author also mentioned to dropout layer after hidden layer to save model from over fitting.

3) **Obesity prediction model**

Obesity is connected with lifestyle and health literacy so, the goal of this model building is to use those variables and predict patients BMI (body mass index) will rise within the next 7 days. This paper proposed multiple models to predict obesity.

- random forest was implemented with 100 estimator (100 decision tree), each tree has minimum 2 split and maximum depth of tree is 1.
- Another algorithm they implemented is a decision tree with a minimum of 1 leaf node and minimum 2 split at child node.
- They implement linear discriminant with Lsqr solver and auto shrinkage.
- AdaBoost was implemented using 50 different estimators and set the learning rate to 1.
- DNN was built using two layers fully connected network with softmax activation function at output layer to predict probabilistic output and PreLU activation function has been used with hidden layers. Also they implemented batch normalization techniques to make the training process fast and stable.

4) **Validation and model assessment**

To assess the stability of the model, 3-cross cross validation is employed. Additionally, the evaluation matrix makes use of specificity, sensitivity, accuracy, and precision. Basically in 3-fold cross validation, two folds are used for training and one-fold for testing at a time. Next time another two-fold use for training and one-fold use for testing. The purpose of doing this is to make our model generalize so when model will perform on blind dataset, it still predicts as accurate output as in training. Additionally, the author uses a single validation dataset in this study for testing and hyperparameter tuning.

VI. STATISTICAL ANALYSIS

Here for every type of model, they implemented six different algorithms and after analyzing those algorithms statistically, they choose one best algorithm for every model.

1) **Statistical analysis on pulmonary disease (COPD) prediction model:**

On the testing dataset, decision trees perform poorly for this model with just 79.2% accuracy, whereas DNN performs well with the highest accuracy of 92.1%. On the validation dataset, however, Random Forest outperforms DNN with an accuracy of 80.4%, about 8% higher. Thus, we may conclude that DNN models may overfit during training and underperform on validation datasets. If we can observe only sensitivity and specificity on the model, we observe that we got more specificity than sensitivity which conclude that our model tends to predict more True negative (TN) than true positive (TP). Additionally, we obtained the highest value for specificity and the lowest value for accuracy across all algorithms.

2) **Statistical analysis on panic attack prediction model**

Here With accuracy of 97.5% and 81.3% in the testing and validation datasets, respectively, the random forest outperformed the other models. However, the sensitivity is lower on the testing dataset. This may be a sign of an imbalance in the data since patients who were recruited later had less panic attacks. Surprisingly, DNN performs the lowest in this model, with just 69.4% validation accuracy and 72.2% testing accuracy. When we just examine the model's sensitivity and specificity, we find that we obtained more specificity than sensitivity, suggesting that our model is more likely to predict true negatives (TN) than genuine positives (TP). Furthermore, i observed that all of our algorithms' accuracy decreases in the validation dataset when we compare their accuracy between testing and validation datasets, indicating that all of our models are overly optimistic and do not generalize correctly. We also obtained the highest value for specificity and the lowest value for sensitivity across all models for all performance measures.

3) **Statistical analysis on obesity prediction model:**

Here The decision tree and random forest performed better, achieving 95.3% and 96.7% accuracy, respectively, whereas all other algorithms produced results between 80% and 85% accurate. For DNN, we got almost double sensitivity in validation than in testing which indicates during testing, model was predicting less true positive (TP) value but, in validation test they predict model more true positive value. So, we can conclude that when we use DNN in obesity prediction model, our model pessimistic which mean our model is under-train during training.

VII. FEATURE ENGINEERING

Feature Engineering is an important part of ML model building so, basically when we build any kind of machine learning model, the model is trained based on dataset (also called features). So, to a make model which predicts accurately, we need our model to closely align with features and also all the features we use to train model is supposed to be independent form each other's. In the dataset, some features are more important than others and those features are more closely align with model prediction output so, while building a ML model we need to put more weight on those features. We need to make are model biased on those features. We want our model to learn more from those features [3]. In this paper 'The SHAP' was designed in order to interpret the results of prediction models based on cooperative game theory and function for that SHAP module [4] is following under that.

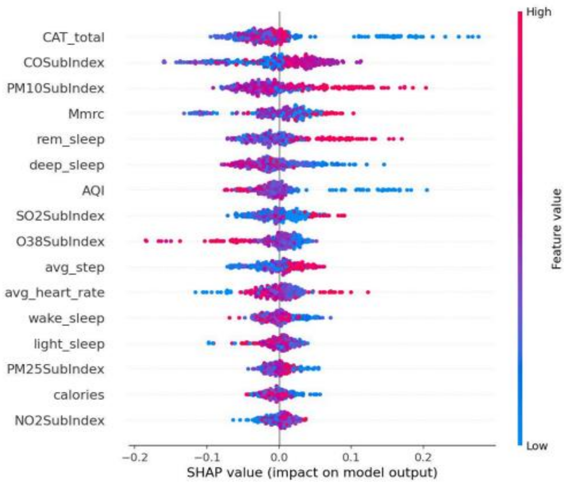
$$\phi_i(p) = \sum_{S\subseteq N/i} x = \frac{|S|!(n-|S|-1)!}{n!}(p(S\cup i)-p(S))$$

Here Ǿ represent shapely value of I feature p(s) is payoff for this collision. S is collision of features. And N is total number of features.

This paper discussed feature engineering for all three model.

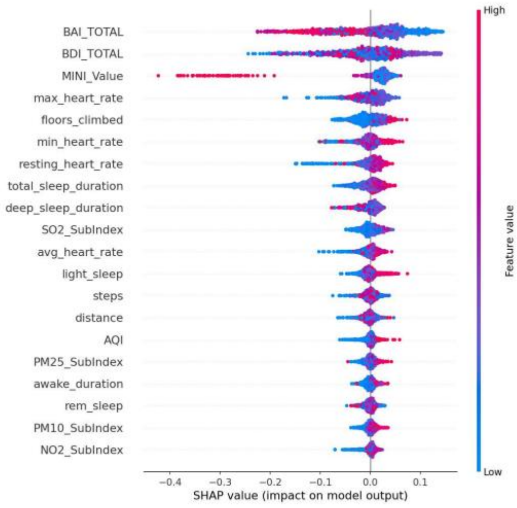
1) **Feature engineering for pulmonary disease (COPD) prediction model:**

In this model we find some features which are more important than others using 'THE SHAP' function. When we train our best algorithm (DNN) on only cost-effective features rather than all features, accuracy on testing was increased by 6%.



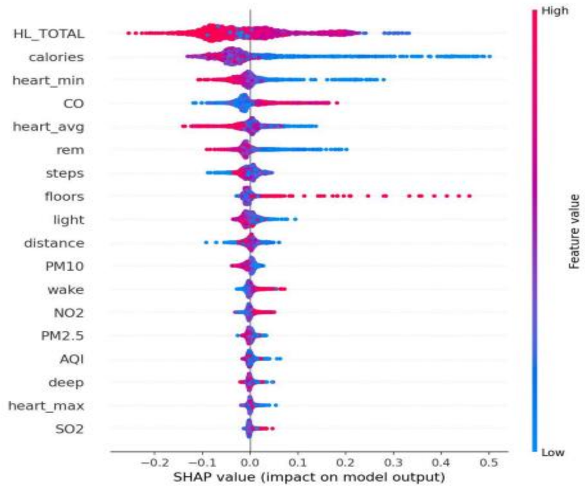
2) **Feature engineering for panic attack prediction model:**

In this phase of model building author has tried to build model by only using important features to check it will impact on accuracy or nor. When 'the shap' module apply on dataset they found BAI_TOTAL, BDI_TOTAL, MINI_VALUE, max_heart_rate and so on are the most important features and when model was train only using those features, accuracy was increased by 2%.



3) **Feature engineering for obesity prediction model**

In obesity prediction model, when 'the shap' module apply on all features and train the model again model achieved an accuracy of 93.7%, a sensitivity of 71.0%, a specificity of 98.1%, and an F1 score of 78.6% on the testing dataset. Also , when we compare models' performance matrix before feature engineering and after features engineering, we found all the performance matrix increased by around 3%.



VIII.APPLICATION IN SOCIETY

- In today's fast-paced world where people has limited time to spend on health, through this AI platform they monitor their health condition and predict the disease in advance and change their lifestyle accordingly to procrastinate those diseases.
- By preventing the progression of chronic diseases, we contribute to the overall improvement of public health within our community.

- The AI platform helps medical practitioners make data-driven decisions, which results in more focused and efficient interventions.
- The early identification of chronic illnesses made possible by the AI-based platform enables preventative medical action and lessens the overall burden on our healthcare system.

IX. CONCLUSION

To conclude, the research showcases a good precision health service for chronic disease prediction, integrating real-time monitoring of lifestyle and environmental factors. This service performs better than earlier research because it offers continuous collecting of data, improves models for predicting chronic diseases, and uses feature engineering to cut down on computing expenses. The model, which serves more than 1,600 patients, reduces the need for repeated hospital visits and provides immediate alerts and decision help. The study highlights the connection between environmental factors, lifestyle choices, and health outcomes, highlighting the value of early warning systems. To establish itself as a leader in the field of precision health, future plans call for increasing data collecting and putting digital twin models for automated health advising into place.

REFERENCES

[1] C. -T. Wu et al., "A Precision Health Service for Chronic Diseases: Development and Cohort Study Using Wearable Device, Machine Learning, and Deep Learning," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-14, 2022, Art no. 2700414, doi: 10.1109/JTEHM.2022.3207825.

[2] Wikipedia contributors. (2023, November 15). *Batch normalization*. Wikipedia, The Free Encyclopedia. **https://en.wikipedia.org/w/index.php?title=Batch_normalization&oldid=1185216804**

[3] J. Heaton, "An empirical analysis of feature engineering for predictive modeling," SoutheastCon 2016, Norfolk, VA, USA, 2016, pp. 1-6, doi: 10.1109/SECON.2016.7506650.

[4] Trevisan, V. (2022, January 17). Using SHAP values to explain how your machine Learning model works. Towards Data Science. https://towardsdatascience.com/using-shap-values-to-explain-how-your-machine-learning-model-works-732b3f40e137

.

.