On Learning from Natural Language to Develop Intelligent System to Identify Potential Medical Condition

Naman Bhoj (✉ namanbhoj99@gmail.com)

Ashutosh Tripathi (✉ at88685@gmail.com)

Research Article

Keywords: artificial intelligence, healthcare AI, Random Forest, Medicine

Posted Date: October 4th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-953323/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

On Learning from Natural Language to Develop Intelligent System to Identify Potential Medical Condition

Naman Bhoj , Ashutosh Tripathi

Email:namanbhoj99@gmail.com, at88685@gmail.com

Abstract: With the rise in human population and emergence of the medical crisis across the whole globe, it has become essential to develop smart systems to automate the process of identification of medical conditions and therefore provide timely aid to the patient. As the first step to develop such a system in this paper we aim to create a natural language based medical condition identification system. The user would provide a text review of how they feel with few other categorical features, based upon that our model would identify the potential medical condition the user is suffering from. We employed three different machine-learning algorithms and mitigated class imbalance in our dataset. Empirical results indicate that Random Forest is the best machine-learning algorithm among all the investigated models with an accuracy of 80.39%, whereas the accuracy of the AdaBoost model improved the highest with an absolute value of 7.77% after mitigating class imbalance.

1 Introduction

According to the latest estimated data there are 7.76 billion people on the planet (U.S. Census, 2021). Managing health related issues of such a huge population is a rising challenge for healthcare facilities and government as in many countries there is a lack of an advanced medical system. It was reported that in 2016, approximately 8.6 million people lost their lives because of the poor medical system across the world which led to diagnosis of medical conditions not being up to satisfactory standard due to the huge volume of patients (M. E. Kruk, et al. 2018). Narrowing down to the human factors which may affect the treatment process of the patient we assume that the experience of a Doctor, decision making capabilities of a Doctor during analysis are main contributing human factors in quality of overall medical diagnosis. Due to the huge volume of patients even experienced Doctors are prone to poor clinical decisions and initial diagnosis of medical conditions, which may result in false positives for a particular condition and consequently leading to wrong medication which may even cause death.

As many diseases spread through direct transmission i.e., if an infected person goes to hospital to get diagnosed, there is a high probability that he can infect other people too. Some common diseases which can spread through contact are Chicken pox, Conjunctivitis, Ringworms etc. apart from this there are some airborne diseases like Mycobacterium Tuberculosis (TB), Common Cold. According to a report TB is considered as the most deadly infectious disease worldwide. Approximately one fourth of the world population is suffering with this disease (W.H.O., 2020). Airborne diseases generally spread when tiny microorganisms of an infected person get mixed with air while he is coughing, talking and sneezing. These diseases can spread even if a healthy person and an infected person are in close proximity in a room or share the same space in a room one after another as these diseases spread through small liquid particles that are ejected from the mouth or nose of an infected person. A person can get infected by intaking these pathogens via mouth, nose and eyes through droplets. These pathogens can stick to the surfaces where they fall upon, therefore touching infected surfaces may cause transmission of these pathogens into the body. Seasonal Influenza and Pertussis viruses spread through droplets. Influenza and pneumonia are the 9th most death causing disease, it causes 15.2 deaths per 100000 patients (CDC, 2019). Hospitals are the hotspots for infectious disease. If a healthy person is going to hospital for medical diagnosis there is a high probability that they may get infected.

Since the start of 2020, the whole world has been going through a very bad situation due to Covid-19. Covid-19 is one of the infectious diseases which is transmitted through droplets, while an infected person is speaking, coughing or sneezing. These droplets can remain on the surface and healthy people can even get infected just by coming in contact with the infected object. On 30 January, 2020 WHO declared it as a Public health emergency. According to the weekly report, 152,534,452 confirmed cases and 3,198,528 deaths were reported worldwide in Weekly Operational Update on COVID-19 on 3rd May 2021 (W.H.O, 2021).

Therefore, to overcome the situation of visiting a hospital in such desperate times for medical condition identification and diagnosis, an automated system which can identify the medical condition in the patients using multimedia data or single stream data like natural language language, image to identify medical conditions is of utmost importance. In the era of technology Artificial Intelligence systems can be trained to perform these classifications very efficiently. All clinical records in text format can be used to train these models.

Figure 1. High-Level architecture of Medical Condition Identification system

As a first step to solve this problem, in our research we aim to devise an automated system which on the basis of single stream text data is capable of identifying the potential medical condition in the patient. The high-level architecture of our system is shown in Figure 1. It shows the architecture of the potential medical condition identification system. The user provides text information which provides details about what the user is feeling and few other categorical variables. The information is then passed to the server which sends the request to the identification system. The identification system generates a response and passes it to the server which serves it to the user. We believe that such a system can be of utmost importance to reduce the spread of various directly transmitted and airborne diseases.

The major contributions made in our paper are:

- Propose a natural language based medical condition system for identifying 25 medical conditions with an average accuracy of 80.389%.
- Discussing the effectiveness of mitigating class imbalance in significantly improving the performance of predictive models.

The rest of the paper is structured as follows:

Section 2 discusses the research conducted in the past, Section 3 discusses our approach, Section 4 analyses the results of our approach and finally Section 5 concludes the paper.

2 Background Research

In this section we go through and analyze the attempts made by researchers in the past to build a system which could classify medical conditions based on a given set of features.

(J. Lui, Z Zhang and N. Razavian, 2018) in their paper proposed models using deep learning techniques such as LSTM, CNN and hierarchical models to predict chronic disease. They used text notes and structured information for classification. Deep learning methods are found much more accurate in comparison to logistic regression and other traditional methods. Their model only predicted 3 classes of diseases namely: heart failure, kidney failure and stroke and attained an average accuracy of: $((90.0 + 83.3 + 75.3)/3) = 82.866\%$.

In their research (H. Yang, I. Spasic, J. A. Keane and Goran Nenadic, 2009) proposed text mining approach for identifying obesity and 15 related comorbidities, therefore identifying 16 different classes of disease. They used a Clinical discharge summary for prediction. They used Textual tasks for predicting disease and intuitive tasks for predicting status of disease. Prediction task is performed on 507 discharge summaries. Their best performing model achieved a F1-Score of 63% for textual tasks.

(S. Mohan , C. Thirumalai; G. Srivastava, 2019) in their paper used machine learning, deep learning and Hybrid Machine Learning techniques for identifying heart diseases. Their proposed method achieved the highest accuracy of 88.7%.

In their paper (G. S. Krishnan and S. Kamath S., 2019) proposed a Feed Forward Neural Network (FFNN) approach for disease prediction. They used electronic Health records instead of manual text notes for prediction tasks of identifying 21 diseases. Their proposed model has achieved the accuracy and F measure of 0.77.

(K. Buchan , M. Filannino and Ö. Uzuner, 2017) in their research they proposed a model that automatically predicts the risk of coronary artery disease. They use free clinical text for prediction tasks. Naïve Bayes, MaxEnt and SVM classification algorithms are used to compare the performance of the model. They used ontology feature extraction methods for optimal results. Their proposed method achieved an F1-Score of 77.4. The highlights of past researchers are aggregated in Table 1.

Table 1. Highlights of Previous Research.

Author	Research Scope	Methodology	Limitation
J Lui, 2018	The research predicts 3 classes	Deep Learning and	The classification was confined to only 3

	of diseases.	Hierarchical Models.	classes of diseases compared to ours which does it for 25 classes.
H. Yang, 2009	The research predicts 16 classes related to obesity and 15 related comorbidities.	They performed text mining over 507 discharge summaries.	The size of the dataset was small as it only contained 507 instances and only 16 classes compared to our 25 class prediction.
S. Mohan, 2019	The research provides direction to classification of cardiovascular diseases.	They used 10 different machine learning and deep learning algorithms.	The research is confined to only identification of cardiovascular diseases while we classify varying diseases.
G. S. Krishnan (2019)	Their research used electronic health records for identifying 21 diseases.	Feed Forward Neural Networks were used.	They identify 21 diseases but use complex Neural Network Architecture.
K Buchan (2017)	Their research used clinical notes to predict coronary artery diseases.	They used Naïve Bayes, MaxEnt and SVM.	The models are only able to classify coronary artery diseases.

Pertaining to the existing research in the literature, we discuss the research conducted in our paper for classifying 25 different diseases in the next section.

3 Proposed Work

3.1 Dataset

For building our model we used an open-source drug review dataset from Kaggle (R.Harode, 2020). Keeping our objective in context i.e., to build a contactless medical condition identification system we treated “condition” as a dependent feature or a variable to be predicted. The detailed data analysis steps are discussed in the next subsection.

3.2 Algorithms

We experiment with 3 different machine learning algorithms for building classifiers for 25-class classification problem.

3.2.1) K-Nearest Neighbour (KNN):

K-Nearest Neighbor is a supervised machine learning technique (L. Peterson, 2009). Which is generally used for both classification and regression problems. It classifies a given data point based on its neighbour data point. Where K is a parameter which decides how many nearest data points should be considered for classification. Given data point is assigned to the class label of the nearest data points. Accuracy of classification increases as well as noise decreases as we consider the high value of K. Lowering the value of K can lead to misclassification. KNN algorithm calculates the euclidean distance from the nearest data point from a given data point. Euclidean distance between two points (x_1, y_1) and (x_2, y_2) , irrespective of dimensions, can be calculated as shown in Eq. (1):

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

For a given value of K, data point assigned with that class, in which K-nearest neighbours are labelled.

Though implementation of KNN is simple, it is a lazy learner algorithm. Instead of learning at time of training, it calculates Euclidean distance for each data point at the time of classification. In each iteration it has to go through the data set, which causes high computational cost of this algorithm.

3.2.2) Random Forest (RF):

Random forest is a supervised machine learning algorithm which is used for both classification and regression problems (L. Breiman, 2001). It uses an ensemble bagging approach for problem solving. It combines results of multiple individual classifiers to get optimal classification. Random forest uses multiple decision trees and to get final results, select the majority of the results given by multiple decision trees. Here random term denotes random selection of data samples for each decision tree and forest term indicates combination of multiple trees. Though inputs at each decision tree may be repeated but repetition may cause decrease in accuracy. In the case of a single decision tree, if we replace a small part of sample data, then we may get large differences in the actual prediction but while performing classification on random forest if we change small parts of sample data, then we may not get large differences in prediction. To achieve better prediction results, Gini impurity and Entropy is calculated at each node of decision trees to ensure the optimal split on each node. The mathematical formula for the same are shown in Eq. (2) and (3)

$$\text{Gini Impurity} = 1 - \sum_{i=1}^N (P_i)^2 \quad (2)$$

Where N represents the number of Classes and p_i represents the probability of class over given input. Gini impurity calculates the probability of misclassification of a class on a node.

$$\text{Entropy} = \sum_{i=1}^N -P_i * \log_2(P_i) \quad (3)$$

Where N represents no. of classes in which input data is being classified and Entropy is calculated for optimal classification of classes on each node. When a sample data is given to the root node of the decision tree, gini impurity denotes chances of misclassification of sample data. Entropy is calculated to ensure optimal classification at each node. Its value lies between 0 and 1.

3.2.3) AdaBoost (AB):

Adaboost is a supervised machine learning algorithm (Y. Freund, and R. E. Schapire, 1996), it uses an ensemble boosting technique for classification tasks. It is also known as an adaptive boosting algorithm because it uses an iterative method to classify a dataset. Adaboost targets the data which are initially misclassified by the base learner. Adaboost classifiers combine multiple poor classifiers to get better accuracy of classification. Adaboost sets weight to each classifier and trains sample data to achieve accurate prediction. By assigning more weight to misclassified data, it increases the probability of classification in the next iteration. Classifiers are also assigned with weight, more accurate classifiers get high weight.

Adaboost is easy to implement and there is a low risk of overfitting with this algorithm. Many base learners can be used in adaboost. It iteratively improves the results of weak learners and combines them to achieve high accuracy.

Adaboost is more sensitive to noisy data. Since it tries to fit perfectly, it may be highly affected by outliers.

3.3 Data Analysis

3.3.1) Imbalanced Data Classification

In this section we discuss the analysis workflow of our research. The dataset initially consisted of 12 features out of which there were 5 numerical, 1 natural language based and 6 categorical features having a total of 362806 instances. We then began to model our problem as a multi-class classification problem. The goal was to build an identification system which could identify the 25 classes of medical conditions based on the patient's explanation of how they feel through text and other features.

EXAMPLE:

PATIENT TEXT REVIEW: "Haven't gotten pregnant so it does it's job. I switched to this brand from another generic. I get nauseous and generally feel tired from this, also more headaches."

IDENTIFIED MEDICAL CONDITION: Birth Control

Therefore, we kept only the top 25 classes of medical condition in our dataset and removed the rest. We then dropped the "Age", "Date", "Drug", "Sides", "Sex" features from our dataset as "Age", "Drug", "Sex" features are clear indicators of few medical conditions and we aim not to ask the user of our medical condition identification system to provide this information as this would also mean storing user data and introduce security concerns. "Date" and "Sides" features were dropped as they didn't seem to provide a lot of information to build a robust natural language based medical condition identification system.

Then null values were removed from the dataset. Following which we were left with 4 categorical independent features namely: {Drug Id, Ease of Use, Effectiveness, Satisfaction, Useful Count}, 1 textual feature namely: {Reviews}. We relabelled {Ease of Use, Satisfaction} features representing them as a negative sentiment for values below 3, neutral for a value of 3 and positive sentiment for a value greater than 3.

This was followed by preprocessing the textual feature. This included the following steps:

Figure 3. Imbalanced Dataset

This huge imbalance could create bias towards the majority classes during training and as a consequence will result in poor classification results of models for minority classes. To mitigate this problem induced by class imbalance, we used Synthetic Minority Oversampling Technique (SMOTE) (N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, 2002). We used it to generate equal instances of all classes. It resulted in majority class downsampling and minority class oversampling to generate equal numbers of instances of each class. This helped us generate an equal number of instances of each class which can be seen from Figure 4. It was found during testing the model that each class now had 5037 instances and hence mitigated the problem of class imbalance.

Figure 4. SMOTE Balanced Dataset

In the next section we in depth discuss and analyse the results of the approach undertaken in the paper.

4 Result and Analysis

In this section we discuss and analyse the results of our model. For evaluating the performance of our models we use Accuracy and F1-Score as evaluation measures.

Accuracy: In a multiclass classification problem accuracy is defined as the sum of True Positive and Negative divided by the number of instances.

F1-Score: It is simply the harmonic mean of precision and recall and is indicative of the test's accuracy.

We discuss the results in two phases. The first phase discusses results of our initial imbalance dataset and the second phase compares and contrasts the classification capabilities of models after mitigating the class imbalance in the dataset.

4.1 Phase One Analysis

We found Random Forest to be the best classifier with the highest Accuracy and F1-Score, the same can be seen from Figure 5 and Table 2.

Table 2. Evaluation Metric for Imbalanced Dataset

Algorithm	Accuracy	F1-Score
K-NN	40.07%	38.97
RF	78.43%	77.63
AB	49.27%	47.32

Figure 5. Accuracy and F1-Score for Imbalanced Dataset

From Figure 5 and Table 2, it is evident that the RF model outperformed both K-NN and AB. The superior performance of RF is due to the fact it is less affected by noise and generalizes better as the number of trees increases because the error reaches a limit with increasing number of trees.

4.2 Phase Two Analysis

Pertaining to the problem of identifying the medical condition based on a given set of text reviews. After building the initial model and testing it on the test data to identify 25 classes of medical condition, we found a huge loophole which could largely affect the accuracy of our model i.e. class imbalance in our data. This class affects the predictive accuracy of our model as it may make our model biased towards majority classes. Therefore, our model would make accurate predictions for majority classes but would misclassify the minority classes. To mitigate this problem SMOTE was employed on both the train and test dataset.

Figure 6. Accuracy Comparison of Imbalanced and Balanced Data Models

Figure 7. F1-Score Comparison of Imbalanced and Balanced Data Models

Figure 6 and Figure 7, clearly indicates the improved performance of K-NN, RF and AB after mitigating the class imbalance in the dataset. Pertaining to the improvement in performance of all our models it can be empirically said that mitigating the class imbalance problem improves the performance of our model. It majorly helps our model during the training time by providing an equal number of examples of all the classes and therefore eliminating the bias of the model towards the majority class.

The quantitative results of improvements of all the models after mitigating class imbalance is shown in Table 3. From Table 3, we can see the accuracy and f1-score of models after mitigating the problem of class imbalance. It was found that Random Forest performed the best yet again in balanced class classification achieving the highest accuracy of 80.39% and f1-score of 80.45 for a 25-class classification problem. We can also see from the same table the quantitative improvement in accuracy and f1-score was the highest for the AdaBoost classifier where the accuracy increased by an absolute value of 7.77 and f1-score increased by a value of 9.93. Therefore, indicating the positive impact of mitigating the class imbalance on the classifier's performance.

Table 3. Metrics for Various Models

Metric	K-NN	RF	AB
Imbalanced Class Accuracy	40.07%	78.43%	49.27%
Balanced Class Accuracy	42.21%	80.39%	57.04%
Imbalanced F1-Score	38.97	77.63	47.32
Balanced F1-Score	41.93	80.45	57.25
Increment Accuracy	2.14	1.96	7.77
Increment F1-Score	2.96	2.82	9.93

5 Conclusion

Our research aims to create a natural language based medical condition identification system. The user provides a text review of how they feel and few other categorical features, based upon that our system is capable of identifying 25 most chronic medical conditions. This is essential for the currently, pertaining to the difficulties caused by the Covid-19 pandemic. We experimented with three supervised machine learning algorithms, while also mitigating the problem of class imbalance for multiclass classification.

The model can further be improved by using Deep Learning, Hybrid Deep Learning, and introducing distinct features like the user emotion or image of the infected/affected area other than the text review. This kind of multimedia data would not only help our model generalize to a wider set of features, but would also increase the predictive capabilities of the model.

Conflict of Interest:

The authors declare there is no conflict of interest.

References

U.S. Census, “U.S. and world Population clock”. Available: <https://www.census.gov/popclock/>.

M. E. Kruk, et al. (2018) "Mortality due to low-quality health systems in the universal health coverage era: a systematic analysis of amenable deaths in 137 countries,"*The Lancet*", vol. 392, pp. 2203-2212.

W.H.O., (2020) “Global tuberculosis report (2020)”, *World Health Organization*, Available: <https://www.who.int/publications-detail-redirect/9789240013131>.

CDC, “Underlying cause of Death (1999-2019)”, Available: <https://wonder.cdc.gov/wonder/help/ucd.html>.

W.H.O., (2021)“Weekly operational update On covid-19 - 3 May 2021”, Available: <https://www.who.int/publications/m/item/weekly-operational-update-on-covid-10---3-may-2021>.

- J. Lui, Z Zhang and N. Razavian,(2018) "Deep ehr: Chronic disease prediction using medical notes." *Machine Learning for Healthcare Conference, PMLR*.
- H. Yang, I. Spasic, J. A. Keane and Goran Nenadic,(2009) “ A Text Mining Approach to the Prediction of Disease Status from Clinical Discharge Summaries,” *Journal of the American Medical Informatics Association*, vol. 16, pp. 596–600.
- S. Mohan , C. Thirumalai; G. Srivastava, (2019) "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques" *IEEE Access*, pp. 81542–81554.
- G. S. Krishnan and S. Kamath S. (2019),"Ontology-driven text feature modeling for disease prediction using unstructured radiological notes." *Computación y Sistemas*, Vol. 23, No. 3, pp. 915–922.
- K. Buchan , M. Filannino and Ö. Uzuner, (2017), "Automatic prediction of coronary artery disease from clinical narratives." *Journal of biomedical informatics*, vol. 72, pp. 23-32.
- R. Harode, (2020), “WebMD drug Reviews Dataset.” Available: <https://www.kaggle.com/rohanharode07/webmd-drug-reviews-dataset>
- L. Peterson, (2009),"K-nearest neighbor." *Scholarpedia*, 4.2: 1883.
- L. Breiman, (2001),"Random forests." *Machine learning* ,45.1 : pp. 5-32.
- Y. Freund, and R. E. Schapire, (1996) , "Experiments with a new boosting algorithm." *International Conference on Machine Learning (ICML)*. vol. 96.
- Wordcloud, (2021), Available: <https://pypi.org/project/wordcloud/>.
- K. S. Jones, (1972) ,"A statistical interpretation of term specificity and its application in retrieval." *Journal of Documentation*.
- N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, (2002) "SMOTE: Synthetic Minority Over-Sampling technique," *Journal of Artificial intelligence*, vol.16.