# Group 5: Speech to text processing in Medical field

Madhvikaben Bhatt 1154135 | Khyati Patel 1172720 | Osheen Baby Varghese 1168517 | Dhruvisha Patel 1159961 | Niharika Sojitra 1170232

## Abstract

*NLP approaches and AI models are desperately needed in the medical profession to give better and faster therapies to patients. We came across the speech-to-text processing issue for our project after discovering that there isn't much work done in this area in the medical profession, which could help medical personnel treat patients more effectively. In this study, we attempted to give patient prescriptions in the form of speech, which we subsequently converted to text using NLP text processing techniques and models, which were then used to identify the disease category. By reducing unnecessary physical contact with documentation and form completion, this approach will save doctors' time when writing extensive medical prescriptions and lessen the risk of disease spread. Multinomial DB, Random Forest, SVM, and KNN had testing accuracy of 57.88%, 66.43%, 69.55%, and 65.72%, respectively.*

## 1 Introduction

We learned through our research papers that NLP and AI could have a lot greater impact on the medical industry. In comparison to text processing, text classification, information retrieval, and other related disciplines, speech-to-text processing in the medical profession require less effort. We have been working on this issue to lessen the amount of paperwork that doctors must complete to treat patients. Our work process is depicted in the diagram below. As shown in the figure 1, human speech, i.e., medical prescriptions, is provided as the input for our project. This voice is then transformed into text using existing APIs and libraries and further processed to identify illness symptoms. We are utilizing model predictions to forecast the disease class based on these extracted symptoms. We used the MultinomialNB, Random Forest, KNN, and SVM algorithms to train four models. With a score of 69.55 percent, SVM is the most accurate of the four models.

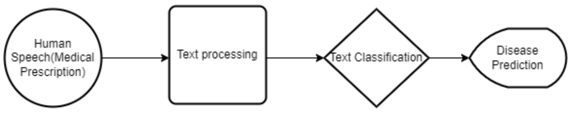


Figure 1: Input and output for the project

To ensure the model's accuracy, we conducted a manual study of 50 medical prescriptions with the assistance of a medical expert. We used this prescription as a set to test all of the models, and SVM came out on top again, with 31 correct illness class predictions. Random Forest, KNN, and MultinomialNB, respectively, have delivered 29, 28, and 26 accurate predictions.

## 2 Motivation

We saw during the COVID pandemic that the medical industry requires considerably greater attention in all of its forms to save human lives. Large amount of the information in the medical area is in the free-text section, which necessitates correct handling. To aid in this effort, we came upon the idea of working on speech-to-text processing, which will allow doctors to minimize physical contact with documents and stationery, reducing the spread of viruses and allowing them to spend more time treating patients. Furthermore, because prior research in this field has primarily focused on text data, it is important to identify a path in this direction as well. Texts in the medical area have a wide range of formatting, sentence structure, and punctuation errors. Atypical language, such as spelling errors, missing expected words, and unexpected POS combinations, are common in medical prescriptions. Working in this field is also difficult due to the wide range of communication techniques and textual subjects.

We focused on speech-to-text conversion with high accuracy, text processing, and language models to receive the related terms and text classification for

Table 1: Class labels and corresponding description of the disease

| Label | Description |
|---|---|
| 1 | Neoplasms |
| 2 | Digestive system |
| 3 | Nervous system diseases |
| 4 | Cardiovascular diseases |
| 5 | General pathological conditions |

our project using NLP technologies. Converting voice into text with great accuracy is also a difficulty for us since it comprises breathing noise and background noise, both of which generate erroneous speech textual forms. Dealing with these challenges, on the other hand, was interesting, and we learnt a lot, which piqued our interest in working on this project.

## 3 Background

The remainder of the paper is laid out as follows. The data gathering procedure is described in Section 3.1. Section 3.2 summarises past work on this subject, its limitations, and our approach to addressing those concerns. The methods for solving the problem, as well as its success rate, will be discussed in Section 4. The results of the analysis and the key findings will be presented in Section 5. The current work constraints and future research prospects will be discussed in Section 6.

### 3.1 Data collection

It is difficult to find medical data in the proper format on the internet. The majority of the datasets we have encountered are already processed and contain numerical data for the columns, which would not work for us because we were looking for textual data to use the NLP techniques and models. However, we discovered a dataset called "Medical Text" with 14438 items on the Kaggle platform Chaitnya(2018). Only two columns, label and medical prescription are present in the dataset. Every prescription in the range of one to five has a "label" connected with it. Table 1 shows our dataset related labels and descriptions.

### 3.2 Related work

Nafiz et. al., believe that structural interventions can provide valid and trustworthy findings with a 92 percent accuracy. They propose a natural language processing-based framework for categorizing diagnostic reports into different specialties. Due to data collection and analytic restrictions, the author's conclusions are biased. They have been unable to test further permutations due to data constraints, resulting in algorithm biases. They also want to strengthen their study by creating a web-based tool based on their conceptual frameworkSadman et al.(2020).

Xieling et. al., used the following prominent methodologies to investigate bibliographical data from Web of Science, PubMed, and Scopus from 2001 to 2018, encompassing performance analysis, science mapping, and text analysis. They used word-by-word examination method for the first time to perform a large bibliometric analysis of the present corpus of knowledge. It was able to overcome the limitations using manual coding or a word statistical method. However, with the use of NLP approaches, there is still a need to improve and ensure the quality of life in clinical trials research Chen et al.(2020).

Viincenza et. al., has worked on the application that concerns the digitalization of medical prescriptions and authorization of health care services. However, to find the personal data, symptoms, pathology, diagnosis and suggested treatments they have used regular expression and string matching techniques which will not serve accurate results because the medical text have different sentence structure and lack of important POS words Carchiolo et al.(2019). Alexandre Trill has followed the same strategy to find out the important words from the prescription for his project Trilla(2009). He has used regular expressions to find out the symptoms ans patient related information from the medical prescription.

A document is viewed as an unordered collection of independent words having one or more occurrences, according to the BoW paradigm. The BoW assumption is employed in several widely used models, including tf-idf and Okapi BM25 based VSMs and Language Models Chowdhury (2010), Zhao and Mao (2017), Tsai (2012). A document is denoted as a numeric vector if a word is represented by a sequence of numeric numbers. The similarity between query and documents is determined by the similarities between query phrases and documents, as the query may be thought of as a combination of termsJelinek(1980), Zhai and Lafferty(2017).

Cohan et al. proposed a model to identify damage events induced by medical errors in patient care and classify them according to their severity levels, which range from a dangerous condition to death, in another paper Sadman et al.(2020). An input layer, an embedding layer, a convolutional layer, a recurrent layer, and an attention mechanism to increase the recurrent layer's performance were all part of the model. However, due to the relative quantity of risk instances relative to the overall number of data in the category, the performance may have been improved even more if some categories were balanced in terms of harm and no-harm cases.

Gunjan Dhole et al.Dhole and Uke(2019) released a paper titled "Medical information extraction using Natural Language Processing interpretation," which focuses on retrieving medical data from narrative clinical records using Natural Language Processing (NLP). It shows how tokenizing, noun entity recognizers, parts of speech taggers, and connection extractors are utilised in NLP to extract medical information from a narrative text. Medical writing, on the other hand, differs from ordinary text in that it includes complicated medical terminologies such as synonyms and disease names. The system suggested in this research attempts to comprehend text linked to a list of symptoms and then provides the appropriate response.

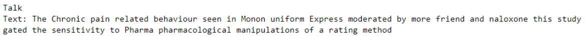
# 4 Methodology

## 4.1 Text Pre-processing

When working with text data, it is self-evident that the data must be cleaned and preprocessed to make it usable for further analysis. This phase was handled with the help of the nltk library. Stopwords have been removed, the text has been changed to lowercase, hyperlinks, punctuation and numbers have been removed. We have corrected the misspelled words using the SpellChecker library[]. Following that, the text was sent to lemmatization, which converted the verbs into their root forms by deleting the *ing, s,* or *es* that were added after the verbs. Lemmatization takes into account the context when converting a word to its meaningful base form, known as a lemma after tokenizing the text using a word tokenizer.

## 4.2 Speech to text conversion

We have used pipwin, pyaudio and SpeechRecognition APIs to achieve this task. These APIs recognize the speech given to the device and generate the text. The figure 2 shows the sample output from the speechrecognizer. Even though there are few words identified by the recognizer incorrectly; it identifies most of the medical words correctly.



```
Talk
Text: The Chronic pain related behaviour seen in Monon uniform Express moderated by more friend and naloxone this study
gated the sensitivity to Pharma pharmacological manipulations of a rating method
```

Figure 2: Speech to text conversion

## 4.3 Vectorization of the text

Machines are incapable of comprehending textual material, which we must work with. So, utilizing the multiple models offered by NLP, we must process the input and transform it into vectors. In a nutshell, we must create vectors of terms and their associated frequencies in the documents. Each medical prescription from the collection is represented by a document. A count vectorizer, a bag of words, and TF-IDF vectorization have all been utilized.

Short for term frequency-inverse document frequency, TF–IDF (also TF*IDF, TFIDF, TF–IDF, or TF–IDF) is a numerical statistic used in information retrieval to express the importance of a word to a document in a collection or corpus. In information retrieval, text mining and user modelling, it is frequently employed as a weighting factor. The TF–IDF value rises in proportion to the number of times a word appears in a document and is offset by the number of documents in the corpus that contain the term, which helps to account for the fact that some words appear more frequently than others in general. The term-weighting system TF–IDF is one of the most prevalent today. We have used all these three models to identify the result variations.

## 4.4 Dataset balancing

When building a model on any dataset, it is critical to balance the datasets. If a dataset is unbalanced, certain classes will have a lot of records while others will have very few. This will cause a problem in model training because the majority of the class label will be used to predict the label for each class. Because our dataset was unbalanced, we employed the random oversampling Brownlee(2015) method to make all of the class records equal. To make the minority class equal to the majority class, this approach creates duplicate dummy records. As shown in the figure 3, records for class two are very less than class five. As shown in the figure 4, after random oversampling all the classes have 2477 records as the fifth class has, the same number of records.

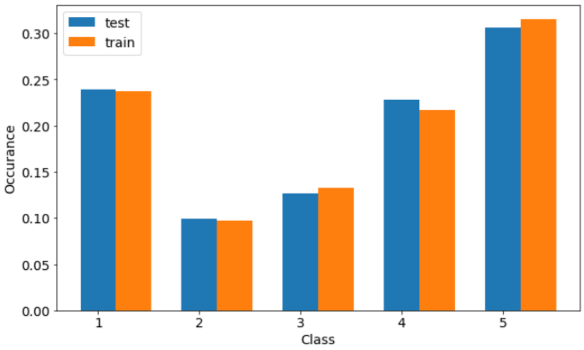
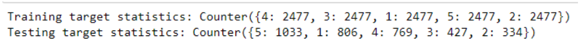


Figure 3: Class imbalance in the dataset



```
Training target statistics: Counter({4: 2477, 3: 2477, 1: 2477, 5: 2477, 2: 2477})
Testing target statistics: Counter({5: 1033, 1: 806, 4: 769, 3: 427, 2: 334})
```

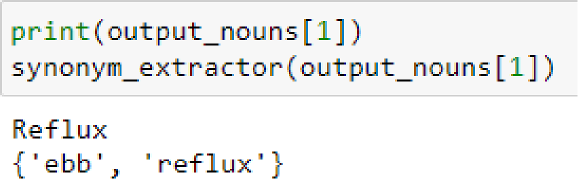
Figure 4: Balanced training dataset

## 4.5 Word similarity check

The usage of a word similarity check is critical in our work because some medical terms are not often used, making it difficult to identify the meaning or make a disease connection. We discovered that terms with the same number of characters are deemed synonyms in earlier research. In other words, two words are deemed similar or synonyms if they share more than three characters from the start. However, this isn't always the case. We used the Wordnet database to locate synonyms for the words to solve this challenge. Because the majority of the symptoms are usually nouns, we gathered nouns from the text and sent them as an argument to the wordnet's synsets function to identify synonyms. We have made all feasible combinations, using these synonyms to identify the prescription's class label. As we can see in the above image 5, all the possible synonyms will be received using the wordnet library.

## 4.6 Model preparation

We utilized the sklearn library for modelling. All of the built-in algorithms are provided as classes in this library, and we only need to import them to use their functions and train our models. To train our models, we employed the MultinomialNB, Random Forest, SVM, and KNN algorithms. For the train and test split, we used a 70:30 ratio. We utilized the ski-kit learn library, which provides the Pipeline class, which does all of this work for us, to use the built-in preprocessing features.



```
print(output_nouns[1])
synonym_extractor(output_nouns[1])
```

```
Reflux
{'ebb', 'reflux'}
```

Figure 5: Synonyms for the word

### 4.6.1 MultinomialNB

For classification using discrete features, the multinomial Naive Bayes classifier Shriram(2021) is appropriate (e.g., word counts for text classification). Integer feature counts are required for the multinomial distribution. Fractional counts like tf-idf, on the other hand, may operate in practice. It predicts a sample's category using the strong Naive Bayes assumption that each feature is independent of the others. As a result, we can calculate the likelihood of each category, and the output will be the category with the highest probability.

$P(A/B) = P(A) * P(B/A)/P(B)$

In comparison to other algorithms, this algorithm is simple to construct and reliable. Furthermore, it is a good choice for text categorization with rapid processing, which is the primary reason for choosing this approach for our model's training.

On the other hand, there are several disadvantages to employing this technique, such as its poorer accuracy compared to other algorithms and its inability to handle numeric data. To investigate the differences between model predictions, we applied four alternative techniques.

### 4.6.2 Random Forest

Random forest Wikipedia(2014) is an ensemble learning method that uses a huge number of decision trees to solve classification, regression, and other problems. For classification tasks, the random forest's output is the class picked by the most trees. This algorithm efficiently analyses vast volumes of data and generates accurate predictions that are simple to comprehend. Because it aggre-

Table 2: Quantitative result

| Model | Training accuracy | Testing accuracy |
|---|---|---|
| Multinomial NaiveBayes | 64.15 | 57.88 |
| Random Forest | 100 | 66.43 |
| SVM | 82.95 | 69.55 |
| KNN | 68.43 | 65.72 |

gates all of the outputs into a single result, it has a high level of accuracy. Over-fitting is to be reduced, resulting in minimal variance and good accuracy. It does, however, have a lot of trees, which makes it complicated and necessitates a lot of computing power and resources.

#### 4.6.3 Support-vector machine

Support-vector machinesWikipedia(2014) are the supervised learning models using learning algorithms that evaluate data for classification and regression analysis in machine learning. This algorithm offers good accuracy and performs faster prediction than the Naive Bayes. When the number of features exceeds the number of samples, this technique is memory economical and adaptable, but it also might lead to over-fitting.

#### 4.6.4 K-Nearest Neighbour

Class membership is the result of k-NN categorization Wikipedia(2014). An object is categorized based on a majority vote of its neighbours, with the object being allocated to the most common class among its k closest neighbours (k is a positive integer, typically small). If k = 1, the object is simply assigned to the nearest neighbour's class. We have selected this algorithm with K=5, as it achieves high accuracy in a wide variety of prediction type problems.

## 5 Results

We examined the model accuracy after text preprocessing and before preprocessing to see if text preprocessing has an impact on the result. Text preprocessing results in 3% rise, according to our findings.

### 5.1 Quantitative Results

SVM had the highest accuracy in testing set with 69.55 percent, as indicated in the table 2. Random Forest ranks second place with 66.43 percent testing accuracy while providing 100 percent accuracy on its training set. The KNN and MultinomialNB models, on the other hand, come in last with 68.43 percent and 57.88 percent, respectively.

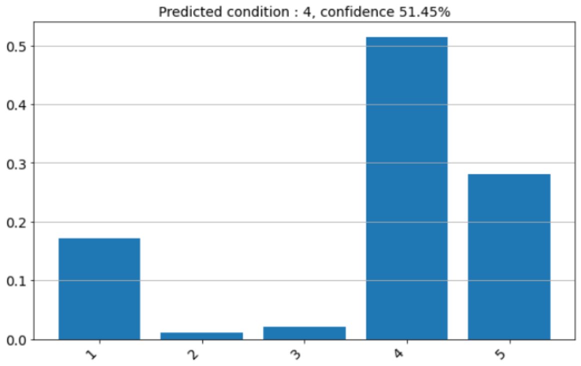


Figure 6: MutinomialNB model prediction for the prescription

The multinomialNB model properly predicted the class of the given medical prescription, as illustrated in the figure 6. The term "confidence" refers to the model's certainty in having that specific label for the supplied prescription. The model is 70.30% certain that the prescription belongs to class four, i.e. cardiovascular disease, as seen in the graph. The Random Forest model has predicted, the class of given prescription as four i.e., cardiovascular disease with 65.50% confidence which is true as shown in the figure 7 below. The predicted result

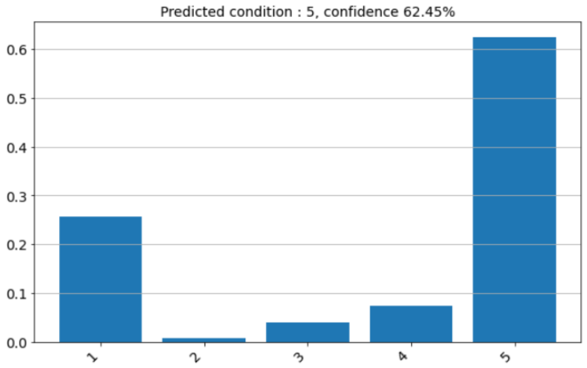


Figure 7: RandomForest model prediction for the prescription

of SVM has shown in the figure 8, which shows that the prescription belongs to class two, which is the digestive system disease with a confidence of 79.80%. As shown in the figure 9, the model has predicted a given prescription as four, i.e., Cardiovascular disease which is predicted wrong with a confidence of 77.40%.

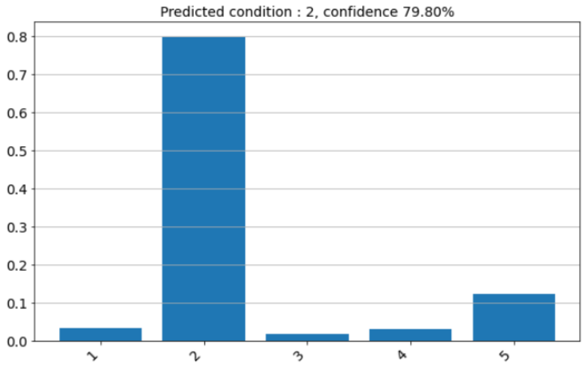


Figure 8: SVM model prediction for the prescription

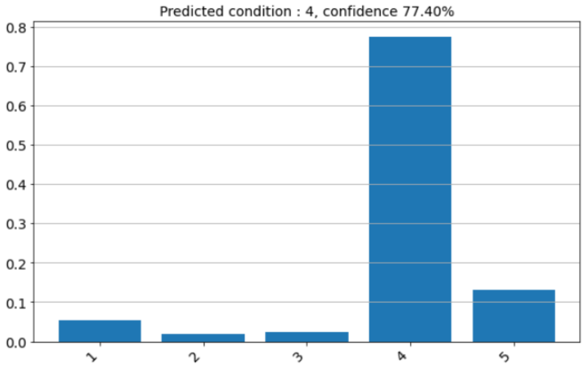


Figure 9: KNN model prediction for the prescription

### 5.2 Qualitative Results

We have performed a manual analysis of 50 randomly selected prescriptions to evaluate the model results. We have allocated labels to each of these prescriptions with the help of a knowledgeable medical field person and also the internet. After training our models, we applied the same 50 prescriptions set to each model and we got the below results. As shown in the table 3, SVM predicts the prescriptions with the highest accuracy. Random Forest, KNN, and MultinomialNB have accuracy in respective orders.
In short, these results are matching with our quantitative study results.

Table 3: Qualitative analysis result

| **Model** | **Result** |
|---|---|
| Multinomial NaiveBayes | 26 |
| Random Forest | 29 |
| SVM | 31 |
| KNN | 28 |

## 6 Limitations and Future work

### 6.1 Limitations

There are some limitations to our work that needs to be addressed. First, after 25-30 words of continuous word listening, the speech recognizer we utilized stops and generates bogus text for the given audio. Second, background noise and breathing sounds were not eliminated from the speech, which could explain why certain uttered words were not accurately detected by the speech recognizer. Furthermore, our algorithms are only trained to forecast five diseases, a number that should be increased by adding new disease categories. Additionally, models must be trained with more records to improve their accuracy, as a single incorrect prediction in this area might result in a significant loss of life.

### 6.2 Future work

For the future paths, we've opted to focus on enhancing voice recognition to correctly forecast speech and eliminate noise []. The medical prescription will be input into the application platform (GUI) and classified. We also wish to focus on the challenge of removing noise from the input speech. Furthermore, we wish to use the LSTM, RNN, and BERT algorithms to train models because they are often used in the articles we have studied for the project literature. In addition, we wish to employ a genetic algorithm to create accurate and efficient forecasts.

## 7 Conclusion

AI and NLP would play a critical role in the medical profession if more study was done. In our experiment, we achieved roughly 70% accuracy for speech to text conversion and 69.55% accuracy for prediction using the SVM model, which is the best of the four models. While Random Forest produced 100% training accuracy, it only provided 66.43% testing accuracy. To complete our duties, we employed a variety of tools and libraries, such as tokenization and language models, which will be useful in our future work projects, as NLP has a promising future in the area of AI and model development.

## References

Jason Brownlee. 2015. Random oversampling and undersampling for imbalanced classification.

Viincenza Carchiolo, Alessandro Longheu, Giuseppa Reitano, and Luca Zagarella. 2019. Medical prescription classification: a nlp-based approach. In *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 605–609. IEEE.

Chaitnya. 2018. Medical text.

Xieling Chen, Haoran Xie, Gary Cheng, Leonard KM Poon, Mingming Leng, and Fu Lee Wang. 2020. Trends and features of the applications of natural language processing techniques for clinical trials text analysis. *Applied Sciences*, 10(6):2157.

Gobinda G Chowdhury. 2010. *Introduction to modern information retrieval*. Facet publishing.

Gunjan Dhole and Nilesh Uke. 2019. Medical information extraction using natural language interpretation.

Frederick Jelinek. 1980. Interpolated estimation of markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice, 1980*.

Nafiz Sadman, Sumaiya Tasneem, Ariful Haque, Md Maminur Islam, Md Manjurul Ahsan, and Kishor Datta Gupta. 2020. “can nlp techniques be utilized as a reliable tool for medical science?”-building a nlp framework to classify medical reports. In *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pages 0159–0166. IEEE.

Shriram. 2021. Multinomial naive bayes explained.

Alexandre Trilla. 2009. Natural language processing techniques in text-to-speech synthesis and automatic speech recognition. *Departament de Tecnologies Media*, pages 1–5.

Chih-Fong Tsai. 2012. Bag-of-words representation in image annotation: A review. *International Scholarly Research Notices*, 2012.

Wikipedia. 2014. Random forest.

Chengxiang Zhai and John Lafferty. 2017. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Forum*, volume 51, pages 268–276. ACM New York, NY, USA.

Rui Zhao and Kezhi Mao. 2017. Fuzzy bag-of-words model for document representation. *IEEE transactions on fuzzy systems*, 26(2):794–804.

## 8 Appendix

### 8.1 Softwares and libraries used

Anaconda (Jupyter notebook): Platform to implement project in python
SpeechRecognition: To recognize given speech
pyaudio: Library to support SpeechRecognition
pandas: Library to deal with dataframe objects
nltk: Library to deal with text processing, and text operations like POS tagging, vectorizing
matplotlib: Library to visualize the results
spellchecke: Library to correct the spellings of the words
sklearn: Library to use inbuilt model functions
imblearn: To balance the dataset
pipwin: Supportive library for the SpeechRecognition

### 8.2 Group Member Contribution

**Madhvikaben Bhatt**
She has done duties for literature review, manual analysis, text pre-processing, algorithm implementation and training models for classification and project related report writings.

**Osheen Baby Varghese**
She has worked on Literature review, data pre-processing and text vectorization.

**Khyati Patel**
She has done Literature review, Identifying limitations of the previous work and qualitative analysis.

**Dhruvisha Patel**
She worked on Speech to text feature, Identifying limitations of previous work.

**Niharika Sojitra**
She has done Speech to text feature,Identifying limitations of previous work.