University of Maryland Baltimore County
Department of Information System

FINAL REPORT

MULTI-LABEL AND MULTI-CLASS CLASSIFICATION OF MEDICAL SPECIALITIES FROM MEDICAL TRANSCRIPTIONS

Dodavah Mowoh

Maryam Alomair

Pronob Kumar Barman

Shreepriya Dogra

IS707: Application of Intelligent Technologies
Dr. Shimei Pan
May 12, 2022

A. ABSTRACT

This paper proposes multi-label and multi-class classifications of clinical transcriptions to predict medical specialties and its efficacy as a decision support tool for medical data classification. In addition, since the data is imbalanced and transcriptions may have more than one label, two transformation methods have been performed to have a more concise data format. SMOTE method was used to oversample the imbalanced labels. The multi-label classification aims to predict the medical specialty(ies) using several classification models, including logistic regression, KNN, Support Vector Classifier, Decision Tree, and Random Forest. Those models have been applied to imbalance labels, balanced labels after applying SMOTE technique, and the One versus Rest classifier. At the end of this experiment, the Multilayer Perceptrons model was used. The experimental outcomes positively demonstrate that SMOTE technique has a significant role in improving the prediction performance. The baseline that applied classification on imbalance labels was brought out better than the One versus Rest classifier. The deep learning algorithm had the poorest performance.

B. INTRODUCTION

PROBLEM STATEMENT

*Multi-label classification of clinical transcriptions to predict the medical codes/*specialties, i.e., the medical departments where the patient needs to be referred.

Need

Medical coding is done by detecting keywords inside medical notes and manually assigning medical codes for each record. This process is time-consuming and suboptimal.

Uses

This study focuses on classifying clinical data that helps healthcare providers distinguish the case based on the transcript. The data can be used for designing solutions for recommending investigations for a thorough diagnosis. There is a chance of identifying medical codes that are not easily identified manually.

Intelligent technologies like ***Text Mining*** and ***Natural Language Processing*** are being used in the project.

MOTIVATION

Medical data is extremely hard to find due to Health Insurance Portability and Accountability Act (HIPAA) privacy regulations. MTSamples dataset offers a solution by providing medical transcription samples. Medical coding is used without machine learning techniques to classify the health record typically performed by a medical coder. The medical coder (tagging staff) detects keywords inside medical notes and manually assigns medical codes for each record to be used by health providers.

MAIN CHALLENGES

The manual coding process is time-consuming and suboptimal. This study focuses on classifying clinical data that helps healthcare providers distinguish the case based on the transcript. The medical data analysis in free-text format provides information such as predicting adverse reactions by analyzing the interaction of drugs (when combined) or identifying co-morbidity risks and forecasting possible conditions that may occur based on previous studies and cases. In addition, the data can be used for designing solutions for recommending investigations for a thorough diagnosis. After pre-processing and classifying the text data, there is a chance of identifying medical codes that are not easily identified. In addition, it may give some insight to tagging staff on how to do the coding optimally.

C. RELATED WORK

Multi-label classification is becoming more popular in text data classification (Schapire and Singer (2000)). Multi-label classification is a supervised machine learning task in which each instance can be assigned numerous labels. **Multi-class**, or **binary classification**, on the other hand, assigns one and only one label to each instance. Problem transformation and adapted algorithms are two extensively used methods for multi-label classification. A multi-label problem is changed into one or more single-label classification problems (Wikipedia (2022a)). To find single-label predictions, various single-label classifiers were used, and subsequently, single-label predictions were turned into multi-label predictions. The method of adapted algorithms has been adapted to make multi-label predictions directly. The techniques of **KNN**, **decision trees**, and **neural networks** are well-known (Wikipedia (2022a)). Each instance is classified into multiple classes in multi-class categorization. Binary transformation, in which multi-class classification is transformed into numerous binary classifications, is a typical approach to multi-class classification (Wikipedia (2022b)). **One vs. rest** and **one vs. one** are two well-known techniques. Extending current binary classifiers to tackle multi-class classification issues is an alternative to binary transformation. **Neural networks**, **decision trees**, **k-nearest neighbors**, **naive Bayes**, **support vector machines**, and extreme learning machines are all popular approaches (Venkatesan and Er (2016)). In classification algorithms, one of the major challenges to dealing with imbalanced data. Data imbalance completely makes the result worse. Data imbalance is becoming a prominent challenge in machine learning classification. Data classifications are not evenly represented in the imbalanced data. Working with imbalanced data is challenging because of the low precision of minority class performance. **Synthetic Minority Oversampling Technique (SMOTE)** is a well-known method for dealing with data imbalances (Chawla et al. (2002) Chawla, Bowyer, Hall, and Kegelmeyer). First, it randomly selects a minority class A instance and finds the first k nearest neighbors. Next, a synthetic instance is created by randomly choosing one of k neighbors B. Join A and B to form a line segment. The synthetic instances are created by combining the two chosen instances, A and B, in a convex way (He and Ma (2013)). **Deep Learning** has gotten a lot of attention in recent years because of its revolutionary work in fields like image classification, speech recognition, and machine translation. **Multilayer perception** is a deep learning algorithm. Multilayer perceptron has at least three layers: an input layer, an output layer, and one or more hidden layers with many neurons stacked together. While neurons in a Perceptron must utilize an activation function like ReLU or sigmoid to enforce a threshold, neurons in a Multilayer Perceptron can use any arbitrary activation function (Carolina(2021)).

D. Dataset Description

MTSamples is a collection of Transcribed Medical Transcription Sample Reports and Examples. It is text medical data containing electronic medical transcripts taken by various transcriptionists. Mtsamples.com gets updated frequently, but our dataset is from 2018. The dataset was scraped from mtsamples.com, which contains **5999 records** of sample transcription reports for many medical specialties and different work types by various transcriptionists and users. The dataset can be used by learning and working medical transcriptionists for their daily transcription needs. It is a public domain dataset. The dataset is imbalanced as various users collected it.

METADATA

Attribute Name	Attribute Data Type	Attribute Description
Number	Integer	Row number
Description	String	Short description of transcription
Medical Specialty	String	Medical specialty classification of transcription
Sample Name	String	Transcription title
Transcription	String	Sample medical transcriptions
Keywords	String	Relevant keywords from transcription

The columns we are using for our project are -

Attribute Name	Attribute Data Type	Attribute Description
Transcription	String	Sample medical transcriptions - long text
Medical Specialty	String	Medical specialty classification of transcription - short text

E. PROPOSED SOLUTION

SYSTEM OVERVIEW

01	**EDA - First Round**	• Exploring the transcription [Text Data] • Exploring medical specialty [The Label]

02	**Pre-processing Data**	• Missing Values & Duplicates
		• Transformation of Data
		  ○ Multi-class Classification Problem
		  ○ Multiple Binary Classification Problem
		• Text Cleaning
		• Feature Extraction (TF-IDF)
		• Feature Selection (PCA)

03	**EDA - Second Round**	• Exploring the preprocessed transcription [Text Data]
04	**Validation Approach**	• 70% - 30% for train - test split

05	**Classification Models**	• Binary classification with
		  ○ Modified OneVsRest (OVR) technique
		• Multi-class classification with
		  ○ Imbalanced Labels
		  ○ SMOTE
		• Algorithm Adaptation Methods
		  ○ Deep Learning using MLP

METHOD DESCRIPTION

The data has 5999 records with some rows that have missing data. This experiment went through several steps before predicting each transcription's medical specialty(ies).
The first step is Exploratory Data Analysis, where pandas, NumPy, and text_data libraries are mainly used.

First: Exploratory Data Analysis (EDA) [First Round]

1. Exploring the transcription [Text Data]
   A. 33 rows with missing values.
   B. 2609 transcriptions are assigned to more than one medical specialty, which means there are 1860 unique transcriptions with 5 medical specialties as MAX and 1 medical specialty as MIN, as it is shown in figure [1]
   C. Two identical duplicated rows
   D. 15162758 words in the combination of all transcriptions.
   E. 2420937 total words in the corpus
   F. 22498 unique words in the corpus
   G. The distribution of the corpus is shown in figure [2]

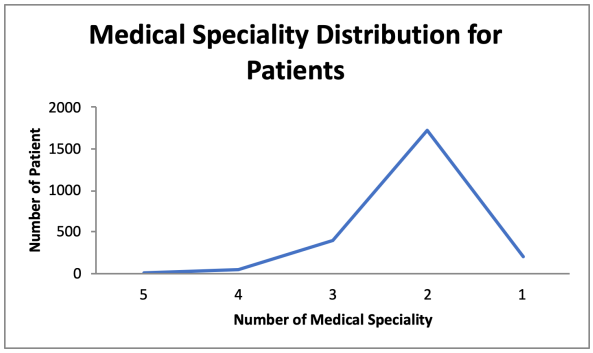

Figure 1

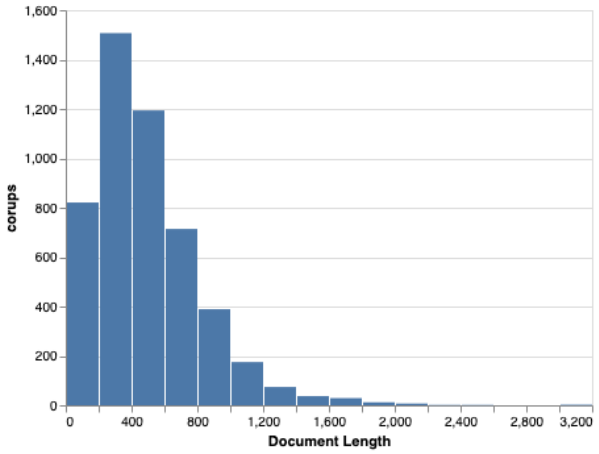

Figure 2

2. Exploring medical specialty [The Label]
   A. 40 classes with 1088 transcriptions/class as MAX and six transcriptions/class as MIN as shown in figure [3]

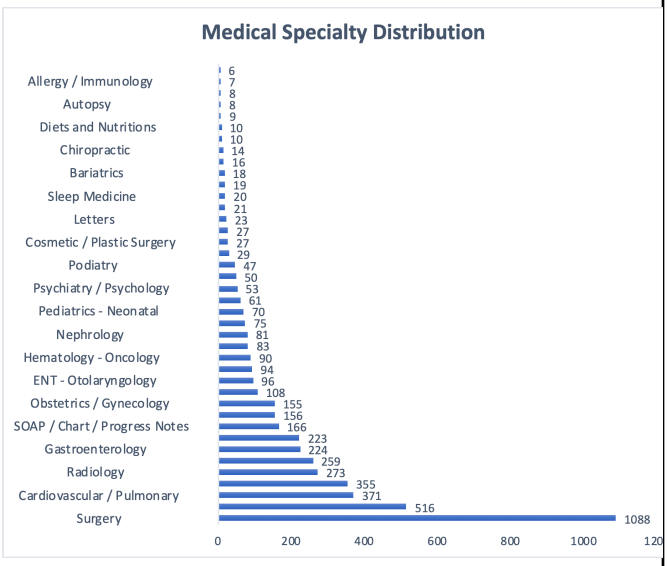

Figure 3

Second: Pre-processing Data

1. Handling missing values: Since the missing data is text and to handle the missing values, the 33 records with missing values have been dropped. The data ended up with 4966 records.

2. Dropping duplicated records: the two identical rows have been dropped. The data ended up with 4964 records.

3. Transformation steps:

A. Transformation into binary classification: Since each transcription has one or up to five labels, there is a need to transform the data to a specific form where each row has all the medical specialties with binary classes, either 0 which means not-assigned, or 1 which means assigned to. To achieve this transformation, a defining Python function is applied to perform the binary relevance method. This step produced 2352 records which consist of transcription and 40 labels. Table [1] shows the original format, as Table [2] shows the transformation.

Transcript	Medical Spec.
D1	M1
D1	M2
D2	M1
D2	M3
D2	M4
D3	M5
D4	M1
D5	M3

Transcript	M1	M2	M3	M4	M5
D1	1	1	0	0	0
D2	1	0	1	1	0
D3	0	0	0	0	1
D4	1	0	0	0	0
D5	0	0	1	0	0

Table 1 & Table 2

B. Transformation into multi-class classification: Another transformation approach has been made where label powerset (LP) transformation is applied for every label combination present in the training set. This step produced 314 combination labels for 1719 records. Table [3] shows the transformation

Transcript	Medical Spec.
D1	M1, M2, M3
D2	M1, M4
D3	M2, M4
D4	M3
D5	M3, M4

Table 3

4. Cleaning text data: since the predictor variable is raw text data. To pursue this step, NLTK and spacy libraries are imported. Several cleaning functions have been applied in a specific sequence as below:
   A. Lowercase convergence
   B. Replace numbers with spaces
   C. Remove stopwords, and medical stopwords
   D. Replace punctuation with spaces except (‘)
   E. Remove stopwords and medical stopwords after removing punctuation marks
   F. Replace (‘) punctuation mark by space

5. Feature Extraction: to pursue this step, the sklearn library is used. Term Frequency-Inverse Document Frequency (TF-IDF) is applied where the importance of term r in a corpus amongst a collection of corpora is calculated, presenting the importance of that term. This vectorization method produced 22461 features (predictors). Since there are many factors, the likelihood of the model overfitting training data and making poor predictions on test data is high. A further pre-processing step is needed.

6. Feature Selection: sklearn library again is used to perform Principal Component Analysis (PCA). PCA is an unsupervised machine learning algorithm widely used to reduce the dimension of data by selecting the most relevant features (vectorized terms in this data) by computing components that explain the most significant variance.

Third: Exploratory Data Analysis (EDA) [Second Round]

After cleaning the data, the text data ended up with:
1. 9684999 words in the combination of all transcriptions
2. 1208089 total words in the corpus
3. 20544 unique words in the corpus
4. The distribution of the pre-processed corpus is shown in Figure [4].

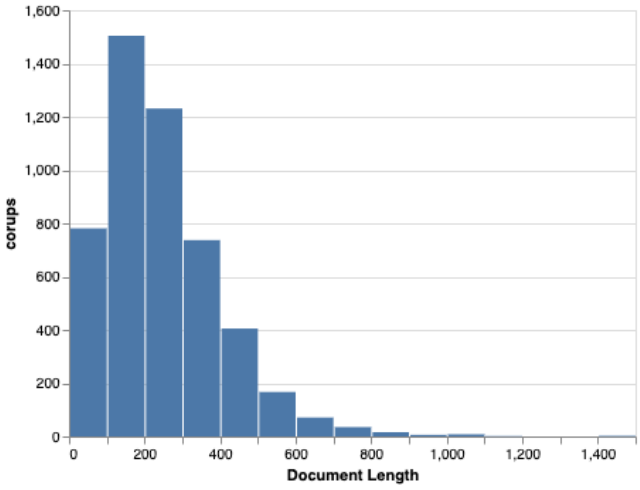

Figure 4

Fourth: Validation approach	
The processed data set is divided into 70% and 30% for train and test sets, respectively.	
Fifth: Classification Models	
Method 1 - Classification models are applied using the data after transforming into binary classification. Those models are under the modified OneVsRest (OVR) technique provided by the sklearn library in Python. A. Logistic Regression B. Support Vector Classifier (SVC) C. Decision tree D. Random Forest E. Ridge classifier	

Method 2 - Multi-class Classification with & without SMOTE

Classification models are applied using the data after applying transformation into multi-class classification through two experiments. To reduce data imbalance among the medical specialties in the dataset two approaches were implemented one with imbalanced distribution for the labels- where all medical specialties with entries less than ten were dropped and SMOTE was not implemented: and a second dataset where Synthetic Minority Oversampling Technique (SMOTE) was applied thus producing a balanced distribution for the labels. The baseline compares the classification results across four algorithms logistic regression, decision trees, support vector models, and k-nearest neighbors.

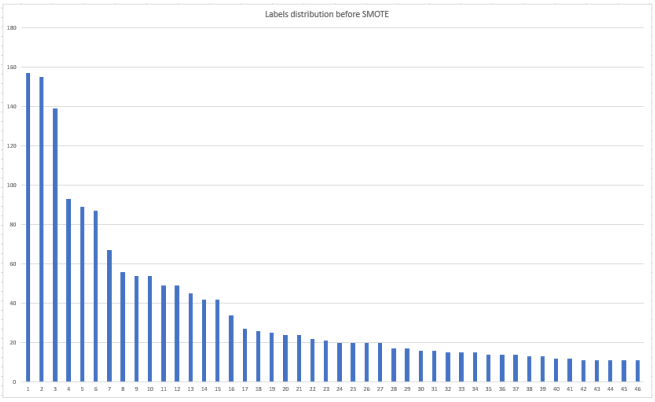

Fig 5

We chose to evaluate the performance of classifiers using logistic regression, decision trees, k-nearest neighbors, and support vector models by using a pre and post SMOTE application dataset. In previous research SMOTE has been considered the “de facto” standard framework for learning about data imbalance (Fernandez et al (2018)) and it

has been proven to be an efficient solution for tackling imbalanced data classification issues(Lui (2022)). Likewise, this approach helps there is one is between-class imbalance, in which case some classes have much more examples than others (Chawla(2004)) in this case the surgery attribute had more examples than majority of the fields in the mtsample dataset.

Additionally, when SMOTE was applied a noisy sample like surgery did not cause the overlapping of minority and majority class which have not yet been properly treated for reducing their influence on the performance of a classification model(Liu(2022)) in the original dataset. Since the synthetic minority samples are generated proportionally to the importance of the minority samples which is calculated according to the composition and distribution of its nearest neighbors (Liu(2022)).

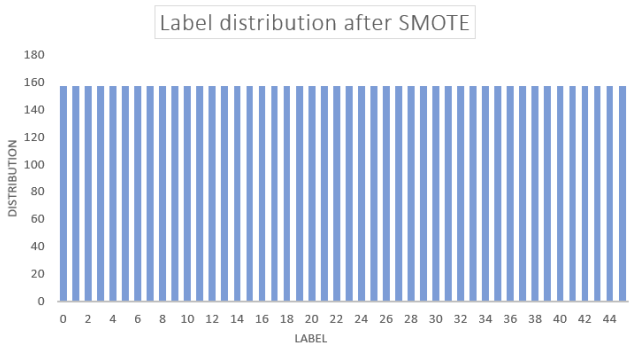

Fig 6

Method 2.1 - Multi-class classification with imbalanced distribution for the labels

A. All labels, which consisted of less than 10 entries, were dropped
B. Figure [5] shows the distribution of the labels.
C. The baseline compares the classification results across four algorithms
    a. Logistic Regression
    b. Decision Trees
    c. Support Vector Classifier (SVC)
    d. k-nearest neighbors (KNN)

Method 2.2 - Multi-class classification with balanced distribution for the labels after applying SMOTE technique

A. Imblearn Python library is used.
B. All the labels have 157 records.

C. Figure [6] shows the distribution of the labels. D. This method compares the classification results across four algorithms a. Logistic Regression b. Decision Trees c. Support Vector Classifier (SCV) d. k-nearest neighbors (KNN)	
Method 3 - Using the same transformed data, deep learning using Multilayer Perceptrons is used A. Keras library in Python. B. linear activation function (ReLU) is used.	

F. EVALUATION

BASELINES

Before running any statistical technique to overcome the imbalanced distribution of the labels, traditional supervised learning algorithms classify the transcriptions based on 314 classes. The classification models are logistic regression, K- nearest neighbor, decision tree, and Support vector machine.

EVALUATION MEASURES

Several metrics have been calculated for fitting the model on the train set and the prediction on the test set. Those metrics are accuracy, precision, recall, F1- score, and hamming loss.

RESULTS

Table [4] Evaluations of training

Model	Accuracy	Precision	Recall	F1 – Score	Hamming Loss
OnevsRest					
SVC	0.99	1.00	1.00	1.00	0.00
Logistic Regression	0.31	0.82	0.52	0.60	0.03
Decision Tree	1.00	1.00	1.00	1.00	0.00
Random Forest	1.00	1.00	1.00	1.00	0.00
Ridge	0.99	1.00	1.00	1.00	0.00

With imbalance Labels [Baseline]					
Logistic Regression	0.81	0.80	0.81	0.77	0.19
KNN	0.75	0.75	0.75	0.74	0.25
Decision Tree	1.00	1.00	1.00	1.00	0.00
SVC	0.62	0.55	0.62	0.57	0.39
With SMOTE					
Logistic Regression	0.99	0.99	0.99	0.99	0.01
KNN	0.80	0.81	0.80	0.80	0.20
Decision Tree	1.00	1.00	1.00	1.00	0.00
SVC	0.99	0.99	0.99	0.99	0.01
Deep Learning					
Multilayer Perceptrons	0.999	1.00	1.00	1.00	0.00

Table [5] Evaluations of testing

Model	**Accuracy**	**Precision**	**Recall**	**F1 – Score**	**Hamming Loss**
OnevsRest					
SVC	0.54	0.89	0.72	0.77	0.02
Logistic Regression	0.24	0.75	0.45	0.52	0.03
Decision Tree	0.24	0.59	0.62	0.60	0.04
Random Forest	0.18	0.71	0.32	0.38	0.04
Ridge	0.50	0.89	0.68	0.74	0.02
With imbalance Labels [Baseline]					
Logistic Regression	0.66	0.61	0.66	0.59	0.34
KNN	0.61	0.60	0.61	0.60	0.39
Decision Tree	0.66	0.68	0.66	0.67	0.34
SVC	0.57	0.50	0.57	0.52	0.43
With SMOTE					
LogisticRegression	0.98	0.98	0.98	0.98	0.02
KNN	0.72	0.73	0.72	0.72	0.28
Decision Tree	0.93	0.93	0.93	0.93	0.07
SVC	0.63	0.63	0.63	0.62	0.37
Deep Learning					

Multilayer Perceptrons	0.02	0.29	0.08	0.10	0.07

Summary of Results

Method 1 - Binary Classification Problems
SVC performed the best with an accuracy of 54% and F1-Score of 0.77.
Method 2- Multi-class Classification - Baseline
Decision Tree & Logistic Regression have similar results. Accuracy for both the models is 66%. Decision Tree has slightly higher F1-Score of 0.67.
Method 2- Multi-class Classification with SMOTE
Logistic Regression performed the best with an accuracy of 98% and F1-Score of 0.98.
Method 3 - Neural Networks
Neural Networks had a very poor performance with only 2% accuracy for the test data.

G. CONCLUSIONS

A considerable amount of research has been conducted to study medical text records that tend to have several patterns. Therefore, mining this type of data helps medical coders to do their work sufficiently. A multi-label classification and multi-class classification methods are used to classify medical transcriptions and assign them to medical specialty(ies). To pursue those two types of classification, two types of transformation were executed to transfer the original data format to two different formats. Both of those formats produce imbalanced labels. Multi-label classification is performed firstly and ends up with acceptable accuracy. Next, Multi-class classification is performed on the imbalanced data and ends up with a similar performance. SMOTE was chosen to oversampled that labeled and built a multi-class classification on balanced labels. This type of classification brought out the highest performance. In the end, to enhance the performance of multi-label classification, MLP was applied. Unfortunately, this model performs poorly and is worse than the traditional classification models.

Evaluation of Multi-class Classification on training data - Significant increase in accuracy scores across all classifiers when models are trained on balance data derived from the application of SMOTE except for decision tree. Models trained on the decision tree algorithm did not change irrespective of the balance or imbalance labels. In other words, accuracy scores were the same at 1.00 for the decision tree classifier pre and post SMOTE. However, significant changes were observed for classifies such as logistic regression, support vector models, and k-nearest neighbors. For logistic regression accuracy for labels with imbalance increased from 0.81 to 0.99 with SMOTE. KNN classifier had a sharp increase in accuracy were the model with imbalance scored 0.75 and model with SMOTE had a score of 0.80. There was major increase in the accuracy score when model was trained using SVC has a score of 0.62 with imbalance labels and 0.99 accuracy score with SMOTE. The best performing classifiers in terms of accuracy for data with imbalance labels were logistic regression and decision tree. With SMOTE decision tree, SVC and logistic regression appear to perform better.

In terms of validation significant changes were observed for Logistic Regression accuracy scores increased from 0.66 to 0.98 for labels with imbalance and labels with SMOTE. Additionally, there was an increase in accuracy for the logistic regression classifier where the score increased from 0.66 with label imbalance to 0.98. Accuracy scores for KNN and SVC classifiers seem to increase only by a small margin.

Further work could be done to study this data. First, since multi-class classification after SMOTE technique was applied gives the best performance in this experiment, SMOTE technique should be applied to the multi-label classification, also looking for better performance.
Second, MLP classification gives us the worst performance after applying it to multi-label classification with the ReLu function. MLP can be performed on multi-class classification with SMOTE. In addition, other activation functions could be applied, such as Scaled Exponential Linear Unit (SELU) and Exponential Linear Unit (ELU).

Future directions of this work that the data can be used for designing solutions for recommending investigations for a thorough diagnosis. The data has another attribute that has diagnostic tests mentioned. By training our model with the data the model can predict additional helpful tests that may be missed by a human being. Also, there is a chance of identifying medical codes that are not easily identified manually and as a preventative action a patient may be reffered to more departments for a screening.

H. REFERENCES

- Transcribed Medical Transcription Sample Reports and Examples—MTSamples. (n.d.). Retrieved February 23, 2022, from https://www.mtsamples.com/
- Parlak, B., & Uysal, A. K. (2015). Classification of medical documents according to diseases. 2015 23nd Signal Processing and Communications Applications Conference (SIU), 1635–1638. https://doi.org/10.1109/SIU.2015.7130164
- Bhavani, A., & Santhosh Kumar, B. (2021). A Review of State Art of Text Classification Algorithms. 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 1484–1490. https://doi.org/10.1109/ICCMC51019.2021.9418262
- Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "A review of multi-label classification methods." Proceedings of the 2nd ADBIS workshop on data mining and knowledge discovery (ADMKD 2006). 2006.
- Elisseeff, A., & Weston, J. (2001). A kernel method for multi-labelled classification. Advances in neural information processing systems, 14.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.
- He, H., & Ma, Y. (2013). Imbalanced learning: foundations, algorithms, and applications. Wiley-IEEE Press.
- Yoram, S., & Learning, M. (2000). Boostexter: A boosting-based system for text categorization.
- Venkatesan, R., & Er, M. J. (2016). A novel progressive learning technique for multi-class classification. Neurocomputing, 207, 310-321.
- Multi-label classification. (2022). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Multi-label_classification&oldid=1070080633
- Multiclass classification. (2021). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Multiclass_classification&oldid=1023970866
- Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis | by Carolina Bento | Towards Data Science. (n.d.). Retrieved May 11, 2022, from https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
- Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.
- Chawla, N., Japkowicz, N., & Kolcz, A. (2004). Editorial: Special issue on learning from imbalanced data sets, ACM SIGKDD Explor., 6, 1–6.
- Liu, J. (2022). Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. Soft Computing, 26(3), 1141-1163.