Universiteit Leiden
The Netherlands

Opleiding Informatica

Data science in the field of heart disease PLN

Tristan Oosterhoorn

Supervisors:
Anne Dirkson & Suzan Verberne

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

dd/mm/yyyy

Abstract

This thesis is about data science in the field of heart disease PLN. PLN is a disease which is caused by a mutation in a specific gene. Since PLN was only recently discovered, there is no medication available and the available information on the disease is limited. This thesis strives to gather insights about PLN using data science applied to two differ data sources: posts from the website forum of the PLN Foundation and online papers. For the website forum, all known medication terms in the data are labelled and used as a base for a CRF model, which uses this information to search for possible new insights in the data. The online papers are searched in Pubmed, which is a extremely large online database with more than twenty-nine million validated papers, articles, books and literature related to the biomedical field. Qualitative research will be used to evaluate the search results of Pubmed. After applying these two methods, the results will be discussed with a cardiologist. The final conclusions led to a few surprising insights, since there were a few medication terms found which could possibly be valuable for PLN patients.

Contents

1 Introduction **1**
- 1.1 Genetic heart diseases . . . 1
- 1.2 Data science for health . . . 2
- 1.3 Research question . . . 2
- 1.4 Thesis overview . . . 2

2 Background **3**
- 2.1 PLN . . . 3
- 2.2 Related work . . . 4
  - 2.2.1 Related to data science . . . 4
  - 2.2.2 Related to applying analyses to same sources of data . . . 4
  - 2.2.3 Related to heart diseases . . . 5
  - 2.2.4 Recent work . . . 5

3 Methods **6**
- 3.1 Data sources . . . 6
  - 3.1.1 Online papers . . . 6
  - 3.1.2 Website forum . . . 7
- 3.2 Overview of the methods . . . 8
- 3.3 Method for online papers . . . 10
  - 3.3.1 Performing queries . . . 10
- 3.4 Method for website forum . . . 12
  - 3.4.1 Preprocessing . . . 12
  - 3.4.2 Labelling . . . 14
  - 3.4.3 CRF . . . 16

4 Results and evaluation **18**
- 4.1 Results . . . 18
- 4.2 Evaluation . . . 22

5 Conclusions and further research **23**
- 5.1 Conclusions . . . 23
- 5.2 Further research . . . 24

References **26**

1 Introduction

In this section the two main subjects of this thesis are introduced. Thereafter, the research question and the thesis overview are provided.

1.1 Genetic heart diseases

Diseases related to the heart can be caused by genetic disorders. Over the last decades, new technologies made it possible to detect more genetic disorders. This has also resulted in more knowledge about genetic disorders which can cause heart diseases. When there is a genetic disorder which causes heart diseases, various effects are possible. These effects can range from little and slightly harmless palpitations till a major decrease of the heart function, which in the end may cause death.[INS]

In essence, the definition of 'genetic' means there is a chance that the disease will be pass along from a parent to a child, which is called heredity. This is illustrated in the figure below. The chance can be familiar, since there are some heart diseases where the chance of heredity is known. This chance may differ between several types of genetic diseases (sometimes the change is 50 percent, but there are also cases where the chance is 95 percent), but can also be unsure. The genetic aspect makes it difficult to control the diseases, because they are transferred as long as a couple has a child before they pass away. The only way this could be prevented would be by curing genetic disorders, but unfortunately this is not possible yet. However, there are ways to make the effects of diseases caused by genetic disorders greatly reduced, but the genetic disorder itself can not be fixed.

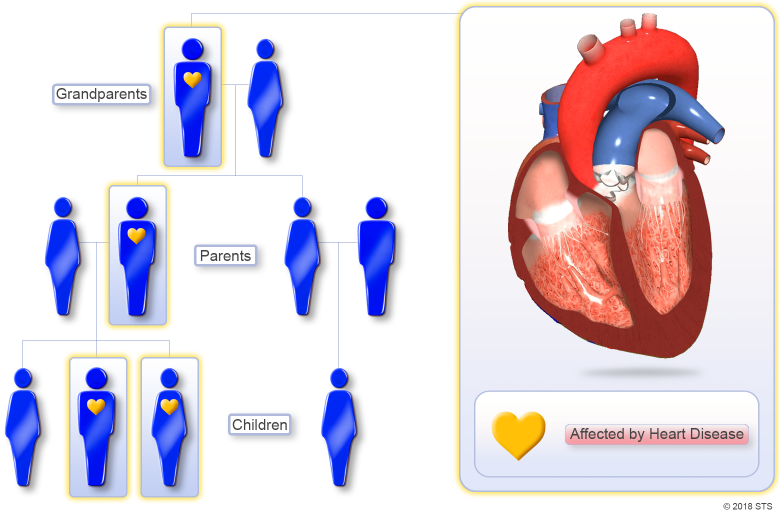

Figure 1: The process of heritable heart diseases. Clarifies the way of spreading genetic heart diseases. Source: www.ctsurgerypatients.org.

1.2 Data science for health

Data science describes the field of research applicable to data, with the goal to obtain insights and knowledge out of the data. This could be done through computational analyses and technical methods. In the last decade, the field of data science has grown a lot in participants and interest of people. This could be dedicated to the growing amount of available data. On one hand, data gets available because of more applications and more people getting access to the internet, and on the other hand storing data gets easier because of better, faster and cheaper data storage systems. [Eco] Because of more data which is coupled with greater use of precision medicine, more personal medical data can be monitored and the data can be valuable for health insights, so that this field is really relevant for data science. In this thesis, data science is specifically applied to texts related to health.

1.3 Research question

This thesis is specifically focused on applying data science to the genetic heart disease PLN. Detailed information about PLN will be provided in section 2.

To understand the essence of the research question, it is helpful to gather some background information about PLN. PLN is a gene in the heart muscle. Mutations in the PLN gene, which are causing heart disease, were discovered in 2012 and since then, a lot of research has been done globally on PLN

However, a medicine for PLN does not exist. Nevertheless, there is a lot of data available on the internet about mutations in the PLN gene. These data are available in literature and the forum of the PLN Foundation website, so this is mainly textual data. Overall, the goal of this project will be to gather insights on medication for PLN, which is not generally known by cardiologists, based on data science techniques applied to the texts which are available in literature and the website forum of the PLN Foundation.

Based on this research goal, the research question is stated as:

Which medication for PLN, which is not yet generally prescribed by cardiologists, can be found with data science techniques applied to the texts which are available in literature and the website forum of the PLN Foundation.

1.4 Thesis overview

After providing an introduction in the main subjects of this bachelor thesis and the research question, the thesis will start by sharing important terminology and definitions which are stated in this thesis. Additionally, related work which has been done on this subject will be shared in section 2. After that, the used method including the results will be described in section 3 and 4. Next, the conclusion and possible future research are described in section 5.

2 Background

Since this thesis is focusing on data about the genetic heart disease PLN, it is useful to understand more about PLN to understand the essence of this thesis as well as possible. Furthermore, to give some background information and to state why this thesis is relevant, some related work will be shared in this section.

2.1 PLN

In order to make our heart beat, calcium ions that flow in and out of the myocardial cells, are very important.

Heart disease is caused by a loss of cardiomyocytes (heart muscle cells), usually because of a heart attack or weakening of the cardiac muscle cells. Heart contractions and the ability to pump blood through our bodies depend on well orchestrated calcium movements in and out of the cell. The PLN (phospholamban) gene is a key regulator of the pump that controls calcium cycling. Scientific names for this PLN gene mutation are c.40_42del AGA and p.Arg14del.

Because of this genetic predisposition, the PLN protein which is read from the DNA, is no longer able to function normally. Mutations in this PLN gene result in enlargement and thickening of heart muscle tissue and arrhythmias, leading to symptoms that include difficulty with breathing, chest pain and fatigue.

PLN cardiomyopathy is a genetic disease with an early onset of symptoms that can lead to sudden premature death. Current therapies, such as lifestyle changes, modern drugs and implanted devices (pacemaker, ICD or Implantable Cardioverter Defibrillator) only alleviate symptoms and prevent the disease from becoming worse. Heart transplant, while not considered a cure, is a life-saving treatment for end-stage heart failure.

Like most inherited heart diseases, the PLN gene mutation itself is rare. As an inherited mutation, PLN cardiomyopathy is known to be common in certain families. Therefor, individuals can be diagnosed and treated early, even before experiencing any symptoms.

Patients already exhibiting disease symptoms, can undergo specific testing for particular disease indications. And, finally, genetic testing of family members of someone with the mutation can become routine care. PLN carriers have a 50% chance of passing this defect onto their children. [Foua]

2.2 Related work

Applying data science to get new insights out of medical and health data is on the rise. This is of course related to the growing amount of available data, as described in paragraph 1.2. Therefor, there is also more and more related work getting available on this subject. In this section, some of these related work will be listed and explained shortly.

This thesis focuses on applying data science to get new insights out of medical and health data of the genetic heart disease PLN. However, methods to apply data science to obtain insights and knowledge which are used for other types of (genetic) diseases can also be useful for the research stated in this thesis.

2.2.1 Related to data science

Currently, there are a lot of research projects where data science is applied to get insights out of medical and health data. One of these relevant research projects is focusing on Alzheimer. The interesting point about this research project, is the way of analyzing the data; this project uses namely the same techniques as used in these thesis, despite the fact that this project is of course much smaller than the Alzheimer related research project.

This related work uses the natural language processing (NLP) techniques to get insights out of the data. In this way, the Alzheimer research project is useful and a inspiration for this thesis; it also confirms that this thesis uses a way of analyzing data which is common in the field of data science.[Xu]

2.2.2 Related to applying analyses to same sources of data

In Jordan, research has been done on diabetes. Despite using different analyses than in this thesis, the research in Jordan has been really relevant for this thesis. Namely, the research used the same sources of data as in this thesis: posts on a online website forum.

Since this thesis is for a large part focusing on data from an online website forum source, it was really useful to get some kind of a confirmation that using data of an online website forum source could be useful and provide insights. [KAS]

Besides the research in Jordan, one of the two supervisors of this thesis also did a research based on the same sources of data. More information about this research can be found in the following reference. [Dir]

2.2.3 Related to heart diseases

Last year, the British Heart Foundation published part of the outcome of a research project where data science was successfully applied to get insights about heart diseases. In this research project, all kinds of data sources were used, from surveys to online papers. The data science process used, among other things, Python to get insights out of the data.

The interesting point of this research project is about the conclusion. This conclusion states that we live in an 'era of digital medicine', which is about getting information about useful and existing medication for specific diseases, in this case heart diseases. In that way, the conclusion of this research project emphasizes one of the assumptions of this thesis: we could be able to get information about useful medication which already exists, based on applying digital techniques on data. [Mit]

2.2.4 Recent work

Recently, in the Netherlands a new research project to lung diseases was published. However this research project uses other data sources and analyzing techniques (including artificial intelligence), the research project confirms that applying data science to get new insights out of medical and health data is a hot topic. [Noo]

3 Methods

This section describes the method of approach for applying data science to get new insights out of data related to PLN.

First, an overview of the data sources is shared. Secondly, an overview of the method is given. Furthermore, the performed method will be explained in detail. Thereafter, the results will be shared in the next section.

3.1 Data sources

Two data sources will be used to perform data analyses and to generate an answer for the research question.

3.1.1 Online papers

The first data source is online papers which are related to PLN.

All over the world, scientists, doctors and scholars publish articles, papers and other literature about the field of health and medication. Estimates indicate that every twenty minutes a new medical paper is published. [Hea].

Some of these papers, are already familiar for the PLN Foundation who gathers information about the PLN gene. The PLN Foundation does not only gather information about the PLN gene, they also coordinate the distribution of information to cardiologists about the PLN gene. However, since there are so many medical papers available, it could be the case that there is much more relevant literature about PLN available in the world.

To research this, Pubmed will be used to perform queries and gather possibly valuable online papers. Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/) is a extremely large online database with more than twenty-nine million validated papers, articles, books and literature related to the biomedical field. All of these data is available as a PDF file trough Pubmed.

The goal will be to find possible valuable online papers via Pubmed, which are not known yet by the PLN Foundation. Namely, the PLN Foundation does not only gather information about the PLN gene, they also coordinate the distribution of information to cardiologists about the PLN gene. In other words, if the online paper is not known yet, the online paper is considered to be a data source which may contain possible valuable information about the heart disease PLN which is not known yet by cardiologists, which is in line with the goal of this thesis.

How this process will be performed is stated in the next paragraphs.

3.1.2 Website forum

The second data source consists of data from a website forum. This forum comes from the website of the PLN foundation, and is meant to share experiences between patients.

This data source can be extremely valuable. Namely, like mentioned in the related work and inspired by the work of supervisor Anne Dirkson[Dir], posts in a website forum can contain information which is not known by cardiologists yet. As an example, people can discuss with each other on a website forum which medication they got prescribed by their cardiologist and what the consequences hereof are.

The reason why this is especially really valuable information, is because PLN is a relatively new and incurable disease, cardiologists sometimes prescribe medication which is not generally accepted for cardiologists. But because PLN is a still incurable, patients get prescribed the 'standard' medication for heart diseases. However, since PLN is a relatively new disease, cardiologists sometimes prescribe medication which is not part of this 'standard' medication.

The above example could be one case where valuable information is stored in the posts of a website forum. To scope the most valuable posts, the method which will be stated in the next paragraphs will mainly focus on posts which contain the names of medication.

An (Dutch) example of a case where a medicine is discussed in a post on the website forum of the PLN foundation, can be seen in the image below, where someone shares a certain medication to use that caused his complaints to be less.

Figure 2: A Dutch example of a post on the website forum of the PLN Foundation. Clarifies it how a post on a website forum can be valuable for getting new insights about a disease. Source: https://hartspierziektepln.nl.

The data is available as a CSV file, which contains 5116 website forum posts. How the process will be performed to make this data useful for analyses and how this analyses method will look like, will be stated in the next paragraphs.

3.2 Overview of the methods

In this paragraph, an overview of the methodology will be stated. The methodology will be explained briefly in text and will be supported by the illustration below (Figure 3) and more specific information in the next paragraphs.

The process will start with collecting the relevant data, based on the two data sources which were explained in the paragraphs 3.1.1 and 3.1.2.
The online papers will be gathered from Pubmed. On this large online database of biomedical literature, several (search)queries can be performed to get an overview of the most relevant literature which is available on Pubmed. How the queries are defined and how the online papers will be selected is stated in the next paragraphs.

The forum data is gathered with the help of the PLN Foundation. They made an export of the database of the website, which is available for this thesis in CSV format. This data contains a lot of noise, as an example website code which is stored in the databases as the posts of the forum. This noise needs to be removed to make sure that the posts of the website forum are useful to perform analyses. The process of making the data 'clean' of noise and useful to perform analyses is called 'preprocessing'. How this process will look like will be stated in paragraph 3.4.1.

After gathering the data and making it ready for analysis, there will be two relevant data sources. On one side a specific amount of possibly useful online papers, and on the other side the data of the website forum containing the posts.

The online papers will be analyzed using qualitative research, in other terms this means that the online papers will be (sometimes briefly) read by the writer of this thesis to determine which online papers can be labelled as a 'potential valuable online paper'. After this selection process the most relevant online papers will be submitted to the evaluation stage of this thesis.

The website forum data will be analyzed using machine learning techniques. During this process, the following steps will be taken:

- Automatically label the posts in the website forum which contains medication names which are already in the standardized list of medication that cardiologists now prescribe to PLN patients. The medication names can either be one of the medications that PLN patients now got prescribed or the brand names of the medication. After this, the outcome will be inspected and reported. More information about this step will be discussed in paragraph 3.4.2.
- Perform CRF training on the data that is the outcome of the above step. More information about this step will be discussed in paragraph 3.4.3.
- Evaluate through cross validation for the known medication names (see step 1) and report possible new medication names found by the CRF model.

At the end of the analyzing process of both data sources, the outcome will be a list of potential valuable online papers and potential valuable medication names from the website forum. This outcome will be reviewed and evaluated. In short, this will be done by comparing the outcome

of the analyses with the list of medication that cardiologists are currently prescribing for PLN patients. The medications that are in the outcome of the analyses, but are not in the existing list will be discussed with a cardiologist. The outcome of this discussion will be stated as the final results in the conclusions section.

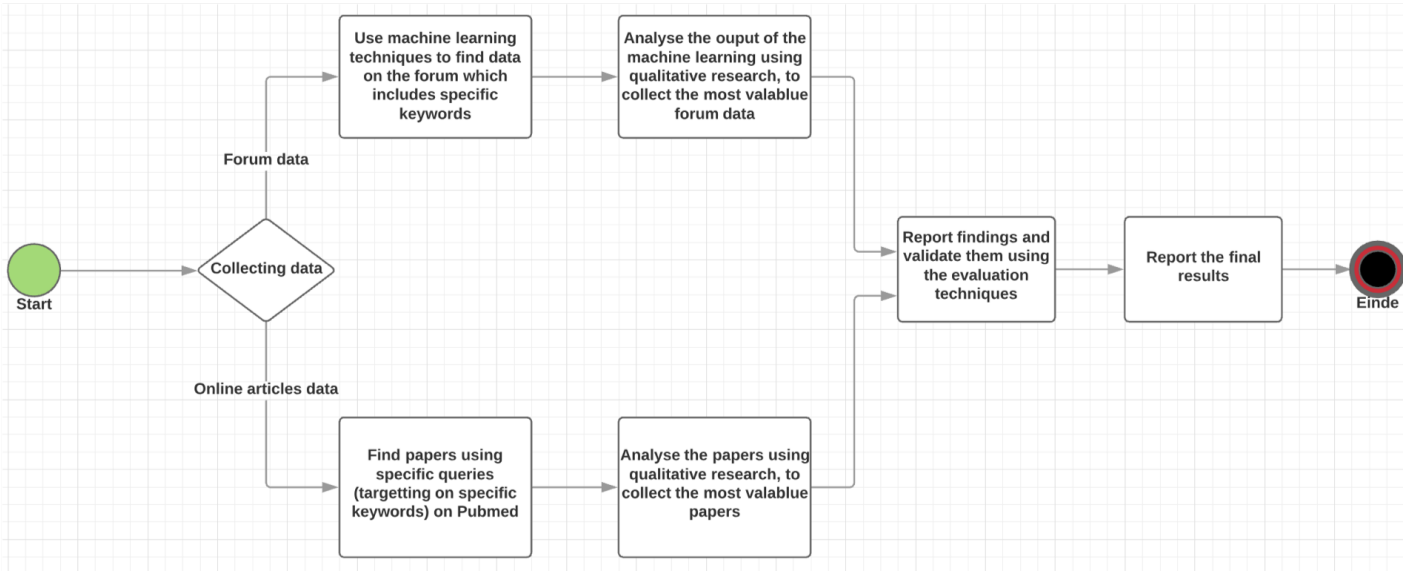

Figure 3: Illustration of the methodology

3.3 Method for online papers

This paragraph will describe the method for selecting and analyses online papers. First, the method to perform (search)queries will be explained in the next paragraph. After that, the way of selecting the online papers which seems to be potential valuable will be stated in paragraph 3.3.2.

3.3.1 Performing queries

As described above, several queries are used on Pubmed to find potential valuable papers which are not known by the PLN Foundation yet. The papers which are familiar for the foundation can be found in the following reference.[Foub] After gathering the papers from Pubmed, they will be quickly analyzed using qualitative research to determine if there are terms for medication used which are not known by cardiologists yet.

The used queries can be find in table 1, including the amount of papers that resulted from this query and are not known yet for the PLN Foundation. The first query was a basic one, since it was only looking for papers which contains the word 'PLN gene' in the text. This query provided three valid papers (valid paper = paper which is about PLN but not known by the PLN Foundation yet).

Having a look at the abstracts of this three papers, it became clear that mostly medical terms are used for PLN. The terms used are R14del and p.Arg14del; the results of this queries are stated in table 1. Additional, the term phospholamban was used a lot, since this is the full name for the abbreviation PLN.

Using the query which searched for papers containing 'phospholamban' as a word in the text gave a lot of results (+100). Most of these papers did not have the disease which is caused by phospholamban as their topic. Therefor, the query needed to be expanded. Firstly, it was tried to search for the word 'phospholamban' in other contexts in the paper, like in the keywords or title. However, this still resulted in too much irrelevant papers. Therefor, the next step was to combine the word 'phospholamban' with other relative terms. One of the words which was combined in a query was 'Dutch'. Namely, PLN is a disease from the Netherlands and was discovered by a Dutch researcher, so therefor it could be logical to think that authors use the word 'Dutch' in their papers to describe the disease PLN. Through this query one valid paper was found. However, this was a paper from a Canadian researcher who was describing that he had several patients with deviations in their heart, and that he had never seen something before. Furthermore in the paper he described that he compared the patient data to data from Dutch patients and in this way he discovered that his Canadian patients are descendants from the Dutch founder of the PLN mutation. More details about this will be desribed in the section 'results'. The other results for queries where phospholamban is combined with a relative term can be found in table 1.

Another interesting view on queries which could potentially generate valid papers, was to search for papers which contains a reference to the paper from the Dutch researcher who discovered PLN. Namely, it could be possible that a researcher (which is not knwon by the PLN Foundation yet) from somewhere over the world was writing about PLN and therefor referencing to the paper about the discovering of PLN. This paper was written by the Dutch researcher Van der Zwaag PA, and

therefor this is used as a query in this thesis. This query generated five relevant papers about PLN.

At this point almost fifty valid papers were found. Since the main focus of this thesis is on analyzing the website forum data, the queries will not be extended in this thesis. However, this could of course be interesting for future research and will therefor be discussed in the section about conclusion and future research. The valid papers have been read (qualitative research) and the results will be discussed in the results paragraph.

If a reader of this thesis is interest in the entire list of valid papers found, please contact the writer of this thesis through t.oosterhoorn@umail.leidenuniv.nl.

Query	Number of valid papers found
PLN Gene	3
R14del	3
Arg14del	9
Phospholamban AND Dutch	1
Phospholamban AND Medicine	9
Phospholamban AND PLN AND Mutation	28
Van der Zwaag PA	5

Table 1: Queries including the numbers of valid papers found on Pubmed

3.4 Method for website forum

3.4.1 Preprocessing

Preprocessing data is extremely important to perform correct and useful analyses. Since the data which is used this thesis is an export of the websites database, the data contain a lot of noise. This could be code of the website, but also irrelevant information like author name of posts, publication date, punctuation marks and many more. All of this data needs to be cleared.

When starting the preprocessing process, the only available data is a large CSV file. To process the data to make sure specific commands and functions can be performed on the data, the programming language Python is used. To convert the data in a way that it is well structured for data analysis through Python, the Python library Pandas is used.

In this thesis, Pandas is used to read the CSV file and to give information about the structure and shape of the file. Through this way, it became clear how the data was structured and how many rows and columns there are in the CSV file. Since this made it clear that the fifth column of the CSV file contained the posts of the website, while the other columns being irrelevant, the following steps in the preprocessing process could be performed in a more specified way. The reason why other columns except the fifth column are irrelevant, is that all of these columns contain data about the technical back-end specifications on the website forum and do not contain any data about the posts on the website forum.

To perform data analysis, the data from the CSV file needs to be loaded in a dataframe. Since it is known that only the fifth column contains relevant data, the fifth column is loaded into a dataframe. When the data is loaded into a dataframe, the noise in the data can be removed and tokens can be created with the support of several filters and normalization. At this point, removing the noise from the data and starting with creating the tokens can be started.

Every post of the website forum will now be 'cleaned', which means that all of the irrelevant items of the data which contains the posts, like website styling code will be removed and the remaining data will be transformed into tokens. HTML and CSS website code, punctuation, stop words and specific characters are removed and processed. While in this process the data gets more clean and the 'actual' words in posts become visible and readable, lemmatization is applied to the data. Lemmatization reduces inflected words properly ensuring that the root word belongs to the language. [Jab]

During the above process, the NLTK package is used. One of the features of NLTK that is used at this moment in the process, is dividing strings into substrings based on the outcome of removing the noise and applying the lemmatization. This substrings are called 'tokens' and are useful in the following processes of applying data science to detect potential valuable information about the heart disease PLN which is not known yet by cardiologists.

The last step during the proprocessing phase is appending the tokens to a list, which is required for labelling the data and to get a valid input for the CRF model. This will be explained in more detail in the next paragraph.

In the following paragraphs, more information is given about how the data, which is now available as the words from the posts from the website forum in tokens (in a list), will be labelled based on the medication terms which are currently known by cardiologists, conform the methodology which is stated in paragraph 3.2.

3.4.2 Labelling

The next step in applying data science on the data, will be labelling the data which contains medication which is already known by cardiologists. The medication that is already known by cardiologists, is the kind of medication that cardiologists now prescribes to patients. It is part of a standardized list of medication which is used by cardiologists in case a patient has a heart disease for which there is no medicine yet. This list was acquired via the PLN Foundation, who had this list in their possession. The list contains the following medication:

- Bisoprolol
- Perindopril
- Dabigatran
- Eplerenon
- Entresto
- Rivaroxaban
- Metoprolol
- Amiodaron
- Sotalol

The medication which is already known by cardiologists is compared with the tokens, which were discussed in the previous paragraph. This kind of ground truth labelling based on partial knowledge is called 'weak supervision'. The medication and the tokens are both placed in two separate Python lists. These two lists are compared using a for loop construction in Python. If one of the medicines is existing in the data, we append a 'B' to the specific word in the list of the data. The comparing process took a lot of time to complete; it could take more than thirty minutes for the computer to run through the entire list of data and compare it with the list of existing medicines.
The final results of this comparing process can be seen in the graph on the next page, figure 4. It is notable that there is one clear outlier in the graph, the Entresto medicine. This seems to be very logical, since Entresto has recently been introduced in the Netherlands. [Nov]
Besides this outlier, it is also intersting to see that most of the medicines are not mentioned on posts on the website forum. Besides Entresto only Bisoprolol and Sotalol are mentioned.

To research if there are more potential interesting words in the data, the brand names of the medicines can also be used as a way to search through the data. Therefor, the supervisor of this thesis provided a list of all brand names which are available on the market for the medicines that are in the standardized list of medication. The outcome of this comparing process is shown in a graph on the next page, figure 5. It is notable that Emcor is an outlier in the results. Emcor is the brand name of Bisoprolol, which is a medicine which is part of the standardized medication list and appears circa fifty times in the data, conform figure 4.

Besides that the outcome as shown in the graphs on the next page specifies specific outliers, the outcome also clarifies that there do exist instances of names of the standardized medicines in the data. This outcome gives confidence for the following steps of this thesis, since it is in the line of the methodology to use existing instances of the standardized medicines in the data to build a machine learning model.

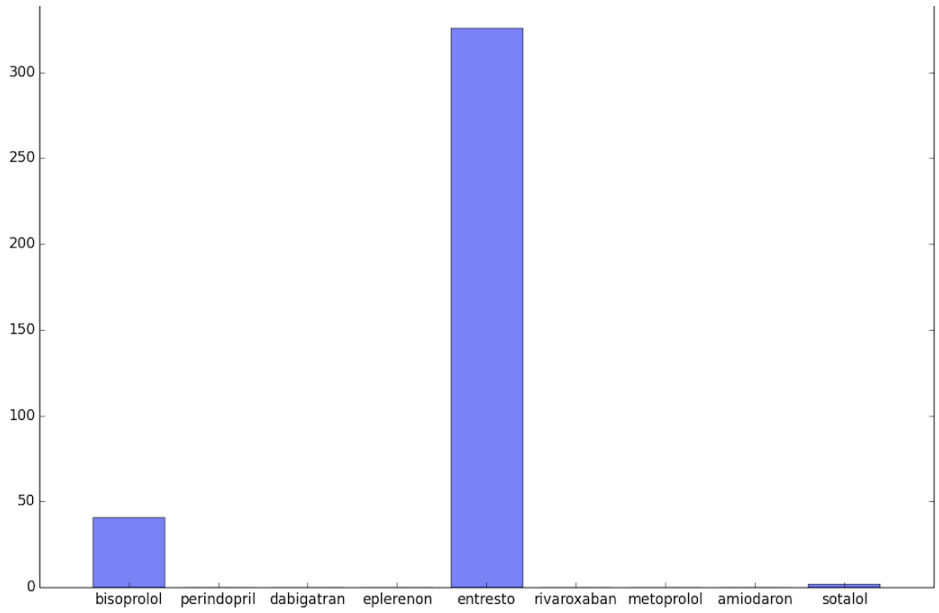

Figure 4: Amount of times one of the known medicines appear in the data. On the vertical axis the amount of times a specific medicine appears in the data is stated. On the horizontal axis the names of the medicines are stated.

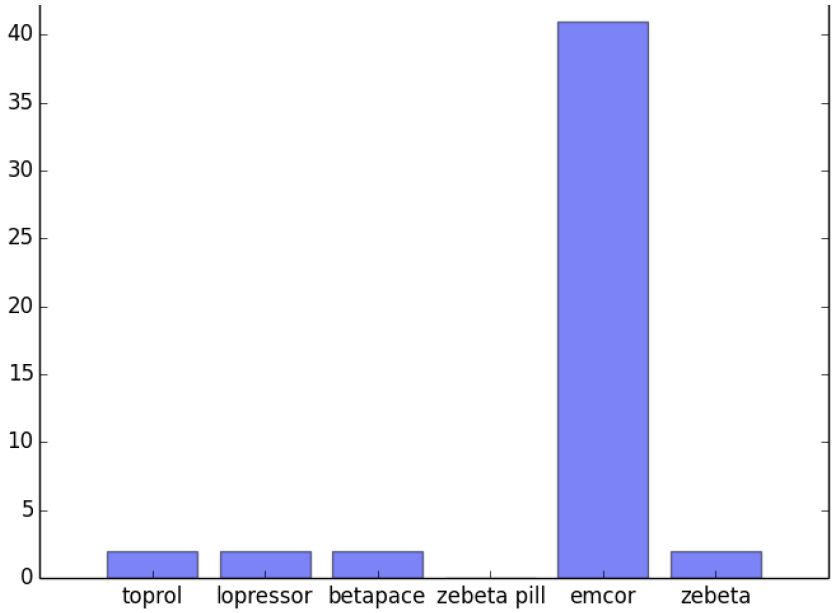

Figure 5: Amount of times one of the brand names for known medicines appear in the data. On the vertical axis the amount of times a specific brand name appears in the data is stated. On the horizontal axis the names of the medicines are stated.

3.4.3 CRF

At this point in the thesis, there is a list which contains all posts of the website forum. Hereby all words which are a medicine or a brand name which do exist in the standardized list of medication has been labelled with a 'B', all other words are labelled with a 'O'. To create a valid input for the CRF model, the entire sentences which do contain a 'B' also get labelled with a 'B'. Furthermore, to start building a CRF model, the data set should be expanded with a Part of Speech (POS) tagging. Namely, in sequence analyzing this type of tagging is needed for a correct analysis and since CRF is a type of sequence analyzing, also this data set should be expanded with POS tagging. [Rac] Therefor the NLTK library in Python, which was earlier discussed in these thesis (paragraph 3.4.1), offers a functionality.

Conform the methodology, the knowledge and structure of the data that has been labelled will be used to perform analyses on the other part of the data. Therefor, a model should be created and trained to be able to find the most relevant posts on the website forum which contains possible valuable information about PLN, which is not generally known by cardiologists.

We use a traditional named entity recognition method for this purpose. Entities are defined as words or sentences in the data that could be interesting for the specific data science goal. In this thesis, the sentences which contains medicines and brand names for the medicines which are part of the standardized list of medication that cardiologists now prescribe for PLN patients are used as entities. These are exactly the instances which are labelled with a 'B', like described in the previous paragraph.

To build a CRF model, the knowledge of the supervisors has been used besides information out of tutorials, such as the official tutorial. [Korb] Hereby specific features are implemented and the data will be split up in a training data set and a test data set.

Features support the CRF model in creating results. This also means that features can improve the result of a CRF model. A first feature was already implemented, the POS tagging, but in this thesis also various features are implemented, namely the identity, word suffix and word shape. This features offer a plain baseline for the model and are suggested by the official documentation as a great standard for features. However, in future research it could be very interesting to experiment with more or less various and differ features, like lower/title/upper flags and features of nearby words. [Kora]

For the split up process various approaches are possible like the 'k-fold cross validation' and 'leave one out validation' approaches. [Bro] In this thesis the default algorithm for splitting is used, the L-BFGS training algorithm. [Hag] In future research, it is recommended to apply the k-fold cross validation to generate a training data set, since this approach is more advanced and guarantees that every observation will be used once for the validation.

After applying the above steps, the model is applied conform the steps of the official documentation. To optimize the outcome, the parameters could be tuned. Hereby hyperparameter optimization is used to improve the quality of the results by selecting regularization parameters using randomized search and 3-fold cross-validation. [Korb] Furthermore, the model will be evaluated through cross

validation and in the end the final classifier is applied to the test data set. The results of the CRF model will be discussed in the following section.

In the large field of Natural Language Processing (NLP), there are several approaches for applying entity recognition. As an example, the Hidden Markov Model and Regular expressions approaches are well known in this field of data science. However, in this thesis the Conditional Random Fields (CRF) approach will be used. The reason for applying this type of approach, is mainly based on the fact of the high rate of successful cases in the world of data science where CRF was applied for entity recognition. [Mac] The choice for CRF is also based on the advice of the supervisors of this thesis.

4 Results and evaluation

In this paragraph, the results and evaluation methods are discussed. First, the results will be discussed in paragraph 4.1. The results will mostly focus on the outcome of the CRF model, since this required a more intensive way of analyzing to get the results. After discussing the results, the evaluation process will be discussed in paragraph 4.2. This will be done by comparing the results with the list of medication that cardiologists are currently prescribing for PLN patients. The medications that are in the results, but are not in the existing list will be stated as the final results in the conclusions section.

4.1 Results

The first thing to evaluate is the precision, recall and the f1-score of the tagger in the scikit-learn classification report. Hereby the precision represents the amount of relevant instances from the retrieved instances, whereby the recall is the total amount of relevant instances that has been actually retrieved from the proportion. To clarify this, see next image (figure 6). Furthermore in the scikit-learn classification report, the f1-score is the harmonic average of the precision and recall. How closer the f1-score is to 1, the better the score is. The last item, the support, represents number of samples of the true response that lie in the class. [slOD]

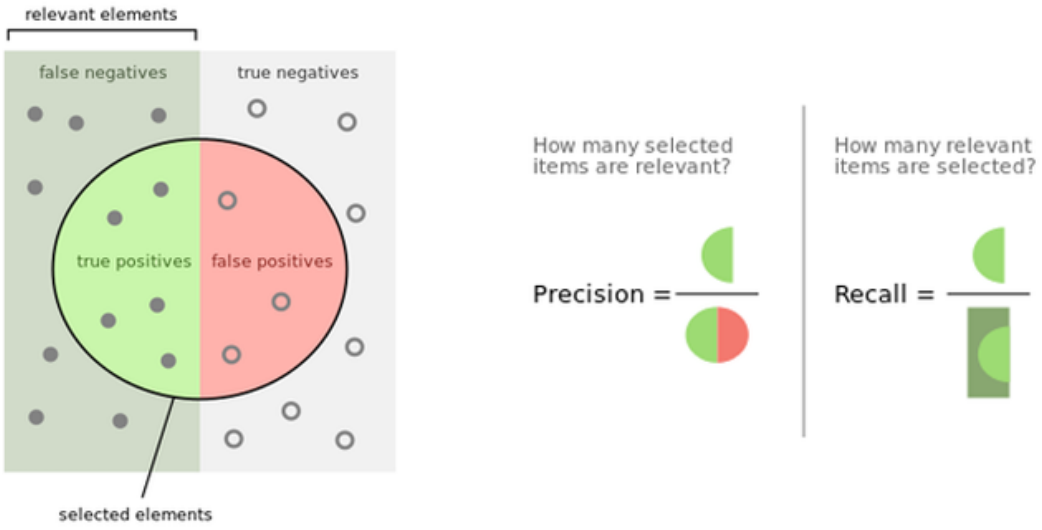

Figure 6: Precision and recall. Source: www.digital-mr.com.

In table 2, the scikit-learn classification report for the CRF model is shown. Based on the information about the precision, recall, f1-score and support which was described in the section above, the f1-score seems to be quite low. However, since the precision is perfect and the recall is low, this should give enough confidence that a model should be able to detect relevant items in the data set which are not selected yet. Namely, the precision shows us that the selected items are all relevant but not all relevant items are select, which also explains the precision and recall. It should be mentioned that the data set is relatively small, which may influence the outcome of the scikit-learn classification report since there is less to analyze. But it provides a basis for CRF models with larger data sets which, will be discussed in the section about the conclusion and future research.

	precision	recall	f1-score	support
B	1.000	0.100	0.182	10
avg	1.000	0.100	0.182	10

Table 2: scikit-learn classification report

After generating the scikit-learn classification report, hyperparameter optimization is used conform the official documentation for CRF models. [Korb] Next, we can have a look at what the classifier learned and which tokens ('words') are stated as results by the model. This can be seen in the image in table 3 and table 4 on the next two pages. Since this thesis only uses a basic implementation of CRF conform the official documentation, more optimization could be done in the future by developing better features and finding better parameters. However, this will be described in the 'future research' paragraph.

Tables 3 and 4 shows some interesting information. Firstly, the word 'middel' is stated as an high top positive word. Having a look in the posts in the website forum where the the word 'middel' is used, it becomes clear that the topic of this posts are mostly related to specific medication. Adding to that, 'middel' can be used as a Dutch word for 'a medicine'. This makes clear that the CRF model is searching in the right place in the data, since we are interest in posts about medication for PLN.

Furthermore, some known names for medications are stated, namely 'entresto', 'bisoprolo' and 'emcor'. However, there is one word that stands out. This specific word is 'diuretica', which is a word that was not known yet and therefor not labelled with a 'B'. Another word that stands out and is not known yet is 'plaspillen', which is the Dutch translation for water pills. Searching on the Dutch website Pharmacotherapeutic Compass, which provides information about registered medications, makes clear that diuretica is a medication name for some kind of water pills. Therefor, there seems to be some connection between the words 'diuretica' and 'water pills'. Since it is interesting to see that the CRF model generated 'diuretica' as a word that could be useful, and diuretica is related to some kind of medication (water pills), this word will be used in the evaluation process in the next paragraph.

Besides the results from analyzing the website forum data, the results from the valid papers which has been obtained through Pubmed should be discussed. Reading these papers gave us the following medication names which could be interesting for cardiologists. These will be evaluated in the next paragraph.

- Pimobendan [Yos]
- Praeruptorin C [Wen]
- Luteolin [Wu1]

Table 3: Top positive:

0.907339 B	-1:word.lower():middel
0.885741 O	BOS
0.821916 O	bias
0.789066 B	word.lower():entresto
0.789066 B	word[-2:]:to
0.789066 B	word[-3:]:sto
0.714423 B	word[-3:]:lol
0.714423 B	word.lower():bisoprolol
0.714423 B	word[-2:]:ol
0.494407 B	-1word.lower():emcor
0.471708 B	word[-3:]:cor
0.471708 B	word[-2:]:or
0.471708 B	+1word.lower():overleden
0.414778 O	word[-2:]:el
0.392558 O	+1:postdag[:2]:VB
0.387551 O	word[-2:]:en
0.376354 B	-1word.lower():naam
0.341844 O	word[-3:]:del
0.341844 O	word.lower():middel
0.321379 B	-1:postag[:2]:RB
0.321379 B	-1:postag:RB
0.294190 O	postag[:2]:VB
0.268826 O	+1word.lower():entresto
0.268252 B	-1:postag:NN
0.249807 O	postag:JJ
0.249807 O	postag[:2]:JJ
0.221892 B	postag:NN
0.210172 B	-1:postag[:2]:NN

Table 4: Top negative:

0.024179 O	word[-2:]:52
0.024179 O	word[-3:]:52
0.021021 O	word.lower():naam
0.021021 O	word[-3:]:aam
0.020585 O	word[-2:]:nd
0.017826 O	+1:postag:VBD
0.013485 O	+1word.lower():eigenlijk
0.006722 O	word.lower():studie
0.006722 O	word[-3:]:die
0.006722 O	-1word.lower():cibis
0.006722 O	+1word.lower():middel
0.000619 O	-1word.lower():begon
0.000619 O	word.lower():diuretica
0.000619 O	word[-3:]:ica
0.000619 O	+1word.lower():plaspillen
0.000619 O	word[-2:]:ica
-0.042087 B	EOS
-0.131865 B	+1:postag:NN
-0.134791 B	word.isdigit()
-0.159617 O	postag[:2]:NN
-0.163188 O	+1word.lower():bisoprolol
-0.165532 O	+1postag:JJ
-0.165532 O	+1postag[:2]:JJ
-0.182150 B	+1postag[:2]:NN
-0.210172 O	+1postag[:2]:NN
-0.221892 O	postag:NN
-0.268252 O	-1:postag:NN
-0.210172 O	+1postag[:2]:NN
-0.221892 O	postag:NN
-0.321379 O	-1:postag[:2]:RB
-0.321379 O	-1:postag:RB
-0.821916 B	bias

4.2 Evaluation

Based on the outcome of the sections above, there are four medication terms which could be interesting.

All of these four terms are not in the standardized list of medication that cardiologists now prescribe to PLN patients. The four medication terms are:

- Diuretica
- Pimobendan
- Praeruptorin C
- Luteolin

Diuretica is a result from the analyses of the website forum data, the other three terms are an result of analyzing the online papers. However, there was not a really connection found between the results of the website forum data en the online papers. This could be due to the small data set, and therefor it could be interesting to expand the research in the future. This will be discussed in the paragraph 'future research'. However, diuretica was discussed in a positively way on the website forum as a patient described it as a medicine that reduces her complaints.

Since all of the four medications are not used by cardiologists yet, because they are not in the standardized list of medication that cardiologists now prescribe to PLN patients, the writer of this thesis discussed the four medication terms with a cardiologist. The cardiologist told us that Pimobendan is currently used as a medicine for animals who are suffering heart issues. However, since the online paper showed that there are some developments going on about using this medicine for humans, this could be interesting for PLN patients and therefor should be researched in more detail.

Praeruptorin C is, according to the cardiologist, a relatively new medicine and currently in the testing phase with animals. Although he was curious about the results, he suggests to wait for the first results of the current testing phase since more judgments about this medication can not be done yet. Furthermore Luteolin is used as a medicine in the Chinese traditional medical world, according to the cardiologist and the online paper. Since this medications are often based on herbs and not on scientific research, the cardiologist is a bit skeptic about this medicine and therefor suggests to not use it in future research to PLN at the moment.

At last, diuretica is a specific kind of a water pill for humans according to the cardiologist. Water pills are currently used for heart patients to make sure there is less moisture in the body. Namely, if there is less moisture in a human body the heart have to work less heart. This is because of the principle that the heart needs to work harder if there is more moisture in the body, because this has to be pumped around in the body by the heart. More information about why water pills have a positive effect on heart patients can be find in the following reference.[Vos] Although water pills are already used for heart patients (and even for PLN patients), diuretica is a specific kind of water pills and not used for PLN patients yet. Therefor it could be interesting to research the benefits of this in more detail, according to the cardiologist.

5 Conclusions and further research

In this paragraph, the main conclusions of this thesis and potential future research are discussed.

5.1 Conclusions

The goal of this thesis was to gather insights on medication for PLN, which is not generally known by cardiologists, based on data science on textual data about PLN. This was formulated in the following research question:

Which medication for PLN, which is not yet generally prescribed by cardiologists, can be found with data science techniques applied to the texts which are available in literature and the website forum of the PLN Foundation?

According to the statements in paragraph 4.2 'Evaluation', there are at this moment two relevant medications which could be useful for PLN. These are Pimobendan and Diuretica. Furthermore, the medication Praeruptorin C is at this moment not relevant since there is too less information about it, but should kept an eye on for the future since it may potentially become more interesting if more research has been done on the medicine. All of this will be shared with the PLN Foundation.

Another insight, which is not specifically based on medication but could be useful for further medical research (since it will expand the amount of patients), is that there are also patients in Canada which seems to have the same kind of heart disease. This has already been communicated with the PLN Foundation and they will distribute this with the research team, since they to the coordination of the research.

Since there has been gathered some new insights about PLN, this thesis can be seen as relatively successful. However, it also became clear that since PLN is a relatively new disease (it is only been discovered a few years ago) and there are not many patients, the data set was quite small. This means that that the CRF model was applied to a relatively small data set and therefor we were not able to train the model a lot. In the beginning of this thesis there was assumed that the data set was relatively large, but during the thesis it became clear that it was relatively small. This does not mean that the CRF model was not successful, since it did generate some useful insights, but in the future it could be more successful if the data set is larger.

At last, the research done in this thesis made it clear that searching online papers for new insights could be highly successful. In this thesis the results from the online paper research may even be called more successful than the results from the CRF model, since the results from the online paper research do have more perspective than the one result of the CRF model, according to the cardiologist. Therefor it can also be stated that using queries on the internet (specifically Pubmed for medical topics) to search for insights, which is not generally known yet, about specific topics could be very useful.

5.2 Further research

The topic of this thesis could be researched in more detail in the future. Although it has been discussed in the previous paragraph that a larger data set could be useful in the future for more research, there are also other ways to optimize the research to PLN with data science.

First of all, like described in the conclusions paragraph, the amount of interesting insights from the online paper research was surprisingly high. Therefor it could be useful the future to do much more research on online papers on the internet. Namely, in this thesis only a few queries were used since the focus of this thesis was on the CRF model. If this queries are expanded and more options are used (like combining more queries, using other meta data to search for in papers) potentially much more insights could be gathered.

Focusing on the CRF model, it could be very interesting to experiment with various differ features, like lower/title/upper flags [Korb] and features of nearby words. [Kora] It is also recommended to apply the k-fold cross validation to generate a training data set, since this approach is more advanced and guarantees that every observation will be used once for the validation. Also with a larger data set, it could be valuable to load more differ data into the model to develop better features and to find the best parameters. This model can then be applied to the original data set and therefor be trained again and again. [Korb] More information about making the CRF model more advanced can be found in the following reference.[Laf]

References

[Bro] Jason Brownlee. Evaluate the performance of machine learning algorithms in python using resampling, 2016. *https://machinelearningmastery.com/evaluate-performance-machine-learning-algorithms-python-using-resampling/.*

[Dir] Anne Dirkson. Knowledge discovery and hypothesis generation from online patient forums: A research proposal, 2019. *http://www.annedirkson.nl/.*

[Eco] The Economist. Data is giving rise to a new economy, 2017. *hhttps://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy.*

[Foua] PLN Foundation. Pln gene, 2015. *https://www.plnheartdiseasefoundation.org/pln-gene/.*

[Foub] PLN Foundation. Pln publications, 2016-2019. *https://hartspierziektepln.nl/publicaties/.*

[Hag] Aria Haghighi. Numerical optimization: Understanding l-bfgs, 2014. *http://aria42.com/blog/2014/12/understanding-lbfgs.*

[Hea] Delve Health. Ghow many medical papers are published each year?, 2017. *https://www.quora.com/How-many-medical-papers-are-published-each-year.*

[INS] UNIVERSITY OF OTTAWA HEART INSTITUTE. Inherited cardiac conditions (genetic disorders), 2019. *https://www.ottawaheart.ca/heart-condition/inherited-cardiac-conditions-genetic-disorders.*

[Jab] Hasa Jabeen. Stemming and lemmatization in python, 2018. *https://www.datacamp.com/community/tutorials/stemming-lemmatization-python.*

[KAS] Omar F Khabour and Ahmed Abu-Siniyeh. Challenges that face the establishment of diabetes biobank in jordan: a qualitative analysis of an online discussion forum, 2019. *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6489903/.*

[Kora] Mikhail Korobov. Named entity recognition using sklearn-crfsuite, 2016-2017. *https://eli5.readthedocs.io/en/latest/tutorials/sklearn$_c$rfsuite.html.*

[Korb] Mikhail Korobov. Tutorial crf, 2015. *https://sklearn-crfsuite.readthedocs.io/en/latest/tutorial.html.*

[Laf] John Lafferty. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 2001.

[Mac] Sidharth Macherla. Complete tutorial on text classification using conditional random fields model (in python), 2018. *https://www.analyticsvidhya.com/blog/2018/08/nlp-guide-conditional-random-fields-text-classification/.*

[Mit] Jennifer Mitchell. Half a million pounds awarded to data scientists researching heart disease, 2018. *https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2018/december/half-a-million-pounds-awarded-to-data-scientists-researching-heart-disease.*

[Noo] RTV Noord. Gronings ziekenhuis zet big data in tegen astma, 2019. *https://nos.nl/artikel/2289836-gronings-ziekenhuis-zet-big-data-in-tegen-astma.html*.

[Nov] Novartis. Entresto zorgt voor medische doorbraak in hartfalenbehandeling, 2016. *https://www.apothekersnieuws.nl/entresto-zorgt-voor-medische-doorbraak-in-hartfalenbehandeling/*.

[Rac] Gianpaul Rachiele. Tokenization and parts of speech(pos) tagging in pythons nltk library, 2018. *https://medium.com/@gianpaul.r/tokenization-and-parts-of-speech-pos-tagging-in-pythons-nltk-library-2d30f70af13b*.

[slOD] scikit-learn Official Documentation. sklearn.metrics.precision$_r ecall_f score_s upport.$, $2007 -$ $2019. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_s$

[Vos] Vitaly V. Vostrikov. Effects of naturally occuring arginine 14 deletion on phospholamban conformational dynamics and membrane interactions, 2015.

[Wen] Wang Wenjiea. Effects of praeruptorin c on blood pressure and expression of phospholamban in spontaneously hypertensive rats, 2014.

[Wu1] Xin Wu1. Erk/pp1a/plb/serca2a and jnk pathways are involved in luteolin-mediated protection of rat hearts and cardiomyocytes following ischemia/reperfusion, 2013.

[Xu] Rong Xu. Big data and advanced artificial intelligence techniques to tackle alzheimer's disease, 2018. *https://www.eurekalert.org/pub$_r$eleases*/$2018 - 11/cwru - bda111218.php$.

[Yos] Oriana Yosta. The r9h phospholamban mutation is associated with highly penetrant dilated cardiomyopathy and sudden death in a spontaneous canine model, 2019.