SMU Data Science Review

Volume 5 | Number 1 Article 7

2021

Machine Learning in the Health Industry: Predicting Congestive Heart Failure and Impactors

Alexandra Norman
Southern Methodist University, abnorman@smu.edu

James Harding
Southern Methodist University, harding@mail.smu.edu

Daria Zhukova
daria.a.zhuk@gmail.com

Machine Learning in the Health Industry: Predicting Congestive Heart Failure and Impactors

Alexandra Norman[1], James Harding[1], Daria Zhukova[2]
[1] Masters of Science in Data Science, Southern Methodist University, Dallas, TX 75275 USA
[2] Lockheed Martin, 1 Lockheed Blvd, Fort Worth, TX 76108 USA

Abstract. – Cardiovascular diseases, Congestive Heart Failure in particular, are a leading cause of deaths worldwide. Congestive Heart Failure has high mortality and morbidity rates. The key to decreasing the morbidity and mortality rates associated with Congestive Heart Failure is determining a method to detect high-risk individuals prior to the development of this often-fatal disease. Providing high-risk individuals with advanced knowledge of risk factors that could potentially lead to Congestive Heart Failure, enhances the likelihood of preventing the disease through implementation of lifestyle changes for healthy living. When dealing with healthcare and patient data, there are restrictions that led to difficulties accessing and obtaining data that have slowed down the ability to apply machine learning and data analytics to problems within the healthcare industry. As data access was limited, this research utilized previous studies and Natural Language Processing to discover common and successful methods to predicting aspects of Congestive Heart Failure. Additionally, this research will look at the most common models used from previous studies and some that have not been applied previously on a common data set used for heart disease analysis. This research will show that the same few types of data or datasets and models have been primarily used over the years and little advancements in analysis for Congestive Heart Failure have been made. Although these methods and datasets have not branched out there is promising results when attempting to predict Congestive Heart Failure on common datasets. With outlining what other data or models could be used this could finally lead to advancements to not only predicting Congestive Heart Failure but utilizing machine learning across the healthcare industry.

1 Introduction

Cardiovascular diseases are the number one global cause of death. There are an estimated 18 million deaths per year due to one of the cardiovascular diseases, with Congestive Heart Failure being at the top of the list [1]. The likelihood of Congestive Heart Failure (CHF) is highly impacted by numerous factors that can be controlled and some that a patient has no control over. Age, gender, and family history are just a few of the factors that cannot be controlled that are known to have a high impact on predicting CHF. Additionally, some personal lifestyle factors and choices, such as smoking, diet, and physical activity, have an impact on the potential to develop CHF. Determining, or predicting, if an individual will have a likelihood of developing CHF,

and the severity of the disease, is related to the presence of risk factors. Understanding the risk factors and early identification for an individual has the potential to decrease the death outcome from CHF. With CHF, prediction is the key to changing the outcome and reducing the mortality rate. Additionally, this study will look to highlight what machine learning models or statistical models have been applied to different data sets over the years. This research will look to uncover the level of impact that each risk factor contributes to the development of CHF. Providing a person with information on their risk factors and probability of onset of the disease may lead to altering behaviors and taking action to reduce the risk and even prevent CHF [2]. Prompt action and intervention through early identification of key factors associated with the disease progression could mitigate the severity of the disease experienced by the individual. Ideally if access to clinical and experimental data was not limited, this or future studies would strive to identify positive lifestyle changes or treatment plans and management plans that could have a positive impact on mitigating the progression of the disease by decreasing the impact of the factors under the control of the individual.

There is a significant amount of people that are living with CHF throughout the world. CHF can affect people of all ages, gender, and race. After developing CHF, the prognosis is that, approximately, half of patients will not survive beyond five years. CHF is one of the leading causes of morbidity and mortality and one of the initial manifestations of cardiovascular disease [3]. Due to CHF having a high morbidity and mortality rate, there has been substantial research conducted to understand the causes and factors that influence someone developing this disease. Research has been conducted across the world on different data pertaining to the information of a person and applying many different models to try to obtain a better understanding of the factors that influence the development and progression of the disease. The goal of many studies is to have a better understanding of factors that lead to developing CHF as this would change the way the healthcare industry handles CHF.

As there is no cure for CHF, it is crucial to detect the early warning signs and potential factors to prevent or manage CHF as early as possible. The more that can be understood about the root causes and warning signs for CHF, the better the heath industry will be able to manage and predict when and how severe the actual incident will be for an individual. With this research, one of the goals is to be able to identify the underlying causes of CHF. Being able to identify and understand those causes will allow for doctors to work with their patients towards making meaningful changes in an individual's life that could greatly reduce the severity of CHF or possibly even prevent the disease from occurring. With the implementation of these changes and the detection of early warning signs, the likelihood of surviving a CHF episode could dramatically increase. Additionally, these changes will decrease the mortality rates of patients already suffering from CHF.

There are numerous factors or combinations of factors that could lead an individual to experiencing CHF. These factors can range from genetical makeup, family history to the lifestyle of the individual, such as diet or exercise. Since these factors will impact both the likelihood of the individual developing the disease, its progression and severity, it is crucial to record them for further research. Many hospitals, clinicals and researchers are already capturing a vast variety of information for the individuals being treated with CHF, however, there is no standard way of capturing the impacting factors. Moreover, the challenge with advanced analytical research is that although a significant

amount of data is captured for patients, most of this data is not accessible to the public. Due to the Health Insurance Portability and Accountability Act (HIPAA), it is not only difficult to obtain patient data unless de-identification processes have been in place but also to correlate and make connections between the different data sets, as individuals' information can exist in multiple research platforms and various databases. In this study, only one readily available dataset was applied to develop predictive models. The study utilized the Cleveland Heart Disease Database which is readily available dataset for heart disease related research and machine learning applications. Since finding and obtaining publicly accessible datasets was challenging, the focus of this study was to also investigate the findings of previously established research. Natural Language Processing was utilized to scrape such information from previously published articles and clinical studies.

The goal of reviewing previous research, methods and analysis using NLP was to highlight the most successful approaches for predicting CHF. This information will provide new researchers with a starting point for potentially new areas of investigation. A concurrent goal of revisiting previous work was to potentially identify new sources of data for developing predictive models on CHF.

For the set of data that was collected, a range of models were evaluated to develop a champion model for providing the best analysis and prediction on the data. In medical research, it is quite common to see more statistical modeling approaches, such as analysis of variance (ANOVA) than the use of Machine Learning applications - Logistic Regression, Random Forest classification model. In this study, statistics were combined with machine learning and Survival analysis to make predictions on CHF.

Upon completion of this research, the ability to quickly understand what has been studied over the years when it comes to CHF as well as how multiple models perform when trying to predict CHF on a common dataset. The first main takeaway from this research highlighted the items that commonly occurred or unutilized across different research studies conducted. One of these items highlighted was the data that had been leveraged when looking to analyze and predict CHF. This piece of information immediately helped indicate what data has been used and what data has been overlooked and not utilized. A driving outcome from calling attention to the data is to discuss the limited access to data and how improving the collection and de-identification process could help enhance data analytics across the healthcare industry. The second piece of information extracted from the previous research studies will show what methods have been utilized to predict various aspects of CHF and if these have improved over the years or if the same concepts seem to be cycling over the years with slight improvement. The final takeaway from the information extracted from the research papers is what have been considered high-risk factors within these different studies and are they all similar. This research took the common models identified in the NLP analysis and some uncommon models and determined which of these performed the best on an extremely common dataset. Comparisons were drawn between what had been previously utilized and what was conducted in this research to see how similar the outcomes are. The outcome of this research will help future studies and the healthcare industry by being a starting point for investigation and to focus on where improvements

can be made with the data and methods used to further explore machine learning for CHF and other topics throughout the industry.

2 Literature Review

Heart failure occurs when the heart's ventricles start working less efficiently than normal causing the heart to pump blood through the body at a much slower rate [4]. There are multiple different kinds of heart diseases that can be led to a person's heart to decrease in its ability to pump blood through the body, these include things like Coronary Artery Disease, heart attacks, Cardiomyopathy, or any number of other conditions that overwork the heart. Overworking the heart can lead to an increase in pressure within the heart's chambers. In turn, the heart responds by either stretching to accommodate more blood in the heart or the chambers become stiff and thickened. These responses by the heart are only temporary fixes to the issue, since overtime the heart's walls will continue to weaken causing the heart to gradually pump less efficiently. Without the heart pumping blood efficiently, the vital oxygen and nutrients are not able to meet the demands of the body. The inefficiency eventually leads to blood and other bodily fluids to backing up into the body's extremities, lungs, or other organs. Once fluid starts building up in the body, it is considered to become congested, and this is referred to as CHF [5].

As with many diseases, CHF can be presented in several different forms or stages. Each stage or form indicates a different level of severity and a general timeline of the progression of this disease. The first stage or earliest form of the condition's progression is often presented with no symptoms during typical physical activities. Management is often done with lifestyle changes, heart medications, monitoring. During stage two, a person may likely experience fatigue, palpitations, or shortness of breath during normal physical activities. As in stage one, the management for this stage can be handled with lifestyle changes, heart medications or monitoring. Diagnosis at either of these two stages along with managing their symptoms and performing the preventative management care may give patients a better long-term outlook. In stage three, a patient is presented with a noticeable limitation of physical activity. During this stage, even mild exercise or completing daily tasks can cause symptoms to be present. Once a person reaches stage three treatment often becomes much more complicated and the life expectancy decreases. The final stage in the progression of CHF is stage four or late-stage CHF. In this final stage a patient generally has symptoms during any kind of physical activity as well as when they are at rest. At this stage, the prognosis for CHF is grim and the likelihood of survival decreases. The longer it takes for diagnosis to be determined for a patient the harder and less likely it becomes to manage and survival with CHF [6].

When it comes to CHF there is thought that there are both factors that can be controlled and factors that cannot be controlled. The factors that can be controlled are things like obesity, smoking, drug and alcohol usage, physical activity, etc. that can

play a role into developing CHF [7]. Since these factors can be controlled these are areas that an individual can look to change to decrease the risk of developing CHF. The factors that cannot be controlled like age, gender, genetics, family history, etc. are areas that an individual cannot change, but these can be used to help determine what kind of treatment or preventative plans will work best.

There have been many studies conducted over the years to get a better understanding of the factors that play into developing CHF and the survival rate or readmissions rate into a hospital [8,9]. These studies have helped the medical industry have a better understanding on what it can expect when dealing with patients that have CHF. The studies that have been conducted vary from predicting which risk factors will reduce the potential of developing CHF to predicting the survivability rate of patients that have different types of heart failure.

A common approach that is used is a statistical analysis approached called the Cox Proportional-Hazard Model. The Cox-Proportional-Hazard Model is a common model used when investigating the effect of several features at the time of a specified event is to happen, like the outcome of death or in this research, the event that a patient is diagnosed with CHF. In previous studies, this approach was used to help with predictions of early detections as well as highlighting which factors can reduced the risk of CHF. In one study, the Cox-based risk prediction model was utilized to identify key factors that affect the prediction of cardiovascular disease [10]. The second study that uses the Cox Proportional-Hazard Model to determine if certain lifestyles that American Heart Association recommend can reduce the potential of developing CHF. This study concluded that with improvement of the lifestyles recommended there was a 55% lower risk associated with CHF [11]. This research will be used to extract and collect which lifestyle recommendation impacted the risk of CHF. This will be one of the treatments and management plan data collected using natural language processing that will then be used to give recommendations of which treatment plans a person should receive. Besides using this model for data collection, this overall concept of using a Cox-Proportional-Hazard Model to predict the likelihood of experiencing CHF will be applied to the different datasets.

Since CHF has been an issue for an exceedingly long time, there have been several different types of methods used to attempt predictions of CHF. These methods range from the more traditional models like logistic regression to newer machine learning models like Gradient-Boosted models. Based on administrative claims data with electronic medical records, machine learning models only show small improvements compared to the more traditional methods [12]. Gradient-Boosted models are one type of decision tree classifiers that will be utilized for this study. Gradient Tree Boosting or Gradient Boosted Decision Trees (GBDT) is a generalization of boosting to arbitrary differentiable loss functions [13]. There are a few deviations of different Decision Trees that were evaluated beyond just the general Decision Tree. Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. Decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules [14]. The Ada Boost

Classifier is a weighted version of a decision tree. The core principle of Ada Boost is to fit a sequence of weak learners (i.e., small decision trees) on repeatedly modified versions of the data. The predictions from all of them are then combined through a weighted majority vote (or sum) to produce the final prediction [15]. Although this previous study shows only small improvements, this research will still evaluate different modeling methods using more than just administrative claims data.

Similar research as to what is being done in this paper has been attempted previously. A study was conducted that uses machine learning to determine the survival rate of CHF based on two factors that the patient is experiencing. Feature ranking was applied to determine the most important risk factors for CHF. This study found that serum creatinine and ejection fraction were the most relevant factors and then built a model using those factors to determine the survival of CHF patients from medical records [16]. The study conducted by Chicco and Jurman seems to focus in on the two factors, when applying machine learning models like Random Forest Classifiers, that were the most relevant based upon the medical records evaluated. Random Forest Classifiers will be utilized in this research, but this research will take into consideration more than just two factors when looking at any survival rate models as well as looking at different datasets. A Random Forest is a collection or ensemble of Classification and Regression Trees (CART) trained on datasets of the same size as training set, called bootstraps, created from a random resampling on the training set itself [17].

An additional piece of data that has been rarely used but could have exponential benefits in furthering the prediction of heart failure is the genetics of the patients. There are markers found within our genetics that could help with the prediction of developing CHF and to what severity. Biomarkers may not just play a role in predicting the likelihood of developing CHF but can be used to help with the decision-making regarding the management of CHF [18].

In more recent years, the importance of utilizing Natural Language Processing (NLP) in the healthcare industry has been increasingly recognized as leading to potential advancements [19]. Natural Language Processing is a subfield within Artificial Intelligence that focuses on languages and giving a machine the ability to understand and interpret what is said or written. This research will utilize the Natural Language Understanding part of NLP as the extraction of information and will not be generating additional text. Although there seems to be more recognition in recent years, NLP within the clinical domain is nothing new and has been utilized since the 1960s [20]. NLP has been utilized to evaluate words, sentences, or document level annotations on numerous types of clinical data like patient status, past medical history, diagnoses, symptoms, treatments, or other information found in electronic health records [19]. In 2014 there was a challenge held by i2b2/UTHealth, in which researchers and practitioners utilized NLP and machine learning to determine different risk factors that led to heart diseases found in electronic medical records [21]. Similar concepts used in the research done with NLP will be applied in this research but used not on specific medical data or records but on research papers found on the web that have been previously published.

Topic modeling is a text-mining application that has been used in the medical and bioinformatics fields. It has been used to detect “instructive” structure in data such as genetic information, images, and networks [22]. Topic modeling is not only a great way of collecting and analyzing large-scale amounts of documents but also visualization of written materials with common topics and offer insight. As there are massive amounts of research done over the years on data analytics with predicting CHF, the utilization of topic modeling will come in handy to quickly gain insights into what has been previously evaluated.

There is a significant amount of data on a patient available to people within the health-care industry from a patient’s medical history, genetics, and real time ECG data. Although most of this data is either anonymized or not all collect for one given person, the most vital answers this research was hoping to give are, "Can one predict what a person’s likelihood of CHF is likely to be based on your demographic, genetic and other data?". Knowing which factors have a higher impact of leading to the likelihood of experiencing CHF can help the medical industry to try preventative treatment if their patients live certain lifestyles or have specific medical indicators. With having a better understanding of how these varied factors affecting the likelihood of CHF, the severity, and genetics, this could lead to further understanding of this disease and to more specialized treatment plans for patients based on factors that can be controlled. Although this research will not be able to fully answer that question due to limited access availability, it serves as a starting point to look at what has been done and what still needs to be done in the health-care industry to being able to achieve the ultimate idea of having specialized treatment or preventative plans that could decrease the amount of CHF patients seen and increase the survival rate of patients that will experience CHF.

3 Data

One of the difficulties that was faced during this research was obtaining data. With no ties to hospitals, clinics or research organizations that would have access to the different data for that would benefit this research. The goal was to utilize many different attributes that could further the research done for predicting CHF and the treatment plans for a person that is experiencing CHF. Due to Health Insurance Portability and Accountability Act (HIPAA) access to the different data and ways to make connections between all the attributes of a person that could impact predicting the likelihood of experiencing or the severity of CHF has been an obstacle that will be addressed by looking at datasets. With the restrictions this research will use common and accessible datasets like Heart Disease Dataset found on the UCI Machine Learning Repository to not as commonly found with treatment plans. Some data will be pulled from CSV files from repositories or TAR files used in other studies while other data will be created by scraping past research papers for additional information like models and factors that lead to CHF.

The first dataset investigated for this paper is one of the most used datasets when looking at heart disease problems. This data is known as the Heart Disease Dataset that was pulled from the UCI Repository. There were 76 attributes within the Heart Disease

datasets. but only 14 will be looked like previous studies only use 14 of these 76 attributes [23]. Within the UCI Machine Learning Repository there are four datasets from across the world, but this initial round of data exploration was done on the Cleveland set which contains 303 instances of patient data.

From the initial work done exploring the data within this dataset, quite a few key indicators were revealed. This data analysis revealed and confirmed that males are more likely to experience CHF. From past studies conducted, the incident rate of CHF is greater than for men compared to women regardless of age [24]. Since the chest pain experienced is subjective and based on a person's pain tolerance there was no direct relation seen with the outcome of having or not experiencing CHF. Although there is no direct relation those that have asymptomatic chest pains do have the most outcomes of experiencing CHF. Resting electrocardiogram (ECG) results showing no direct results but having normal ECG is leads to a positive outcome. Even though it is rare in the data, if you have a ST-T wave abnormality you are three times more likely to have heart disease. Having exercise induced angina is a strong indicator for CHF. Patients are almost three times more likely to experience CHF if they have exercise induced angina. Meanwhile they are less than half the time as likely to experience CHF if they do not have exercise induced angina. The patients that had a flat slope distribution are more likely to experience CHF. Patients that have defected thallium test results also are a strong indicator for if that patient will experience CHF.

After completing the first initial review of the Cleveland dataset, permutation feature importance was implemented. Permutation feature importance is a SciKit Learn library that is used to determine which feature had the biggest impact on predictions. This technique is useful for non-linear or opaque estimators when the data is tabular and highlights which attributes are the most important when looking to do a prediction. The permutation importance is calculated based on a basic Random Forest model that was fitted and it is defined to be the decrease in a model score when a single feature value is randomly shuffled [25]. As shown in Figure 1, the premutation importance indicates that the number of major vessels, exercise induced angina, Stress Test (ST) depression and max heart rate achieved had the biggest impact on predictions.

Weight	Feature
0.0361 ± 0.0564	num_major_vessels
0.0328 ± 0.0359	exercise_induced_angina_yes
0.0328 ± 0.0688	st_depression
0.0262 ± 0.0262	max_heart_rate_achieved
0.0230 ± 0.0161	chest_pain_type_non-anginal pain
0.0164 ± 0.0000	age
0.0164 ± 0.0293	st_slope_upsloping
0.0131 ± 0.0245	thalassemia_reversable defect
0.0098 ± 0.0161	chest_pain_type_atypical angina
0.0066 ± 0.0161	fasting_blood_sugar_lower than 120mg/ml
0.0066 ± 0.0262	thalassemia_fixed defect
0.0033 ± 0.0131	sex_male
0.0033 ± 0.0245	cholesterol
0 ± 0.0000	thalassemia_normal
0 ± 0.0000	rest_ecg_left ventricular hypertrophy
-0.0066 ± 0.0161	rest_ecg_normal
-0.0131 ± 0.0131	chest_pain_type_typical angina
-0.0164 ± 0.0000	resting_blood_pressure
-0.0295 ± 0.0321	st_slope_flat

Figure 1: Permutation feature importance for the Cleveland dataset.

A partial dependency plot was created to evaluate the marginal effect two of the key attributes had on the outcome. The two attributes evaluated were considered indicators in predicting CHF. Figure 2 shows a partial dependency plot of the number of major blood vessels and the maximum heart rate achieved. The plot for the number of major blood vessels revealed that when an increase in the number of blood vessels the probability of experiencing CHF decreases. The plot for the max heart rate achieved is the inverse where when the heart rate increases the probability of experiencing CHF decreases.

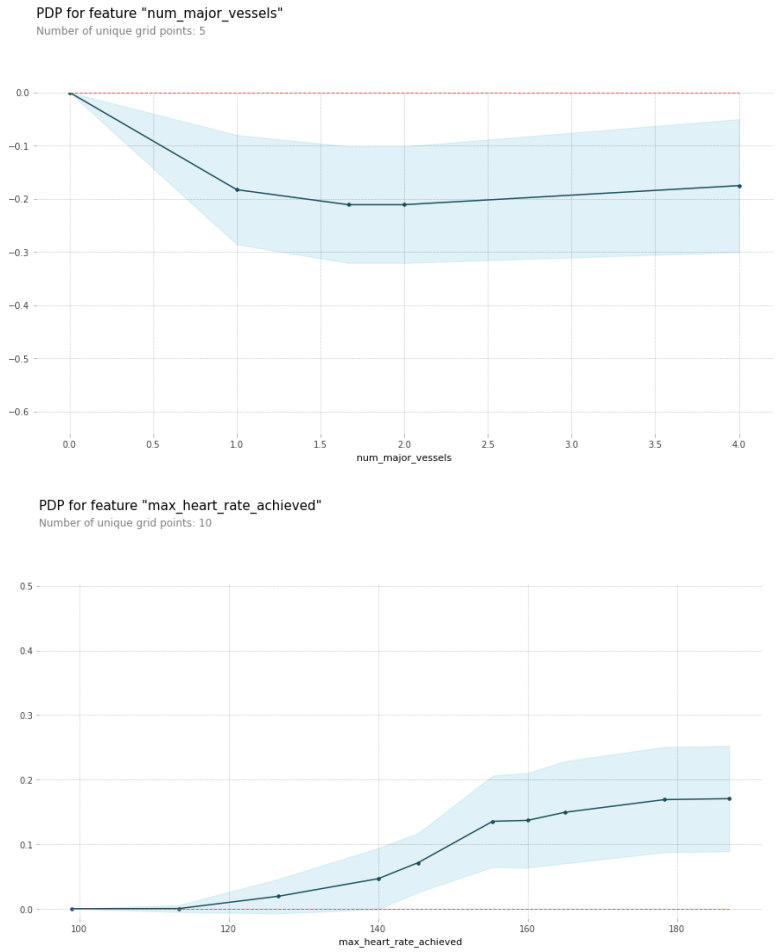

Figure 2: Partial dependency plots for the number of major blood vessels and the maximum heart rate achieved.

From the exploratory data analysis conducted on the UCI dataset, there were a few key takeaways that were revealed to consider when conducting our analysis on

predicting the likelihood that a person will experience CHF. This analysis revealed that a male is more likely to experience CHF than women. Individuals that have experienced heart disease also had a high likelihood of experiencing exercise-induced angina or having an abnormal reading on the thallium test. There were also converse relationships with the likelihood of experiencing CHF and that person have a normal reading on the resting ECG test or would have experienced a fbs <= 120 mg/dl.

Since the Cleveland UCI data set was the only data that was obtained, the set of research papers to evaluate was the next set of data this paper utilized. The collection of these previously completed research studies was pulled off different medical journal websites and consisted of papers that studied data analytics regarding CHF. Currently, one-hundred research papers were collected and the full list of those can be seen in the appendix. These research papers where then stored on a Google cloud drive to be used for consumption within our Natural Language Processing. The first goal for using this large amount of research papers is to get an understanding of what the main topics or ideas were for each of these without having to read numerous pages. A main way to quickly understand key concepts and frequently used words used throughout text is to utilize text mining concepts like topic modeling.

Topic modeling is a commonly used text mining application to determine and understand the abstract topics or frequently used words that occur repeatedly throughout a collection of documents. For the data exploration on the different studies compiled to be evaluated for this research, topic models were utilized to take each document and assign a probability of belonging to a latent theme or "topic." It is important to note that topic models are no substitute for human interpretation of a text— instead, they are a way of making educated guesses about how words cohere into different latent themes by identifying patterns in the way they co-occur within documents. To visually represent these patterns, t-distributed stochastic neighbor embedding (t-SNE) algorithm was utilized to cluster the documents by the dominant topic. t-SNE is an unsupervised learning algorithm that helps with data exploration and visualizing how the data is arranged in high-dimensional space. This algorithm can be used as inputs to a model, but for this research it will be used to investigate and understand the similarities between the different research papers collected. Figure 3 is a plot to visually depict the separation in the data points or research papers. By looking at this plot it is clear to see the distinct groups of data. The selection to create four distinct topics was determined after iterating over a few different numbers of topics ranging from two to twenty and determining that four distinct topics was enough of a distinguishing factor for the papers that were selected to be evaluated. The reason that four topics seem to work is that the research papers that were picked had a narrow look at pulling papers that pertained to machine learning applications for predicting CHF. A wider net could be cast to have more topics to look at but for the goals of this research a more specific collection of research papers works. This visualization could be used to digger deeper into the specific topics or keywords.

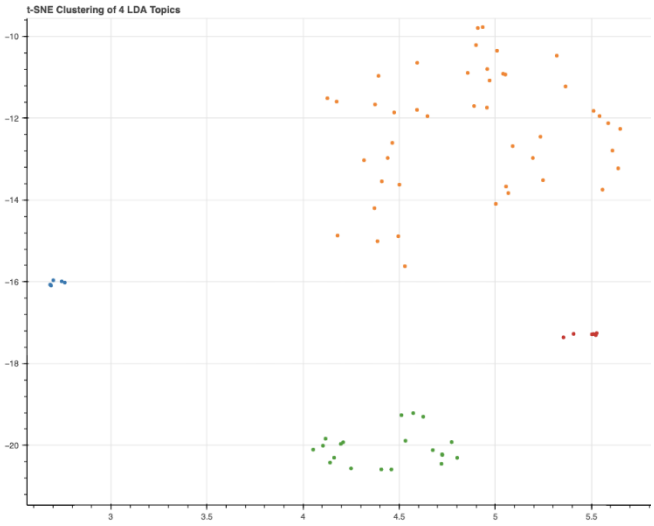

Figure 3: t-SNE clustering of the four LDA topics.

4 Methods

As CHF is nothing new in the medical field, there have been numerous attempts to get a better understanding of root cause and prediction of if someone has a high likelihood of experiencing CHF. These attempts range from statistical modeling to machine learning modeling and even more recently with deep learning methods. The original goal of this research was to apply a wide range of methods covering the entire spectrum of previous models to numerous diverse types of datasets to determine the likelihood of a person experiencing CHF. Since obtaining publicly available data limited the datasets that could be used the focus was shifted to look at previous studies that were done and utilize Natural Language Processing (NLP) concepts to pull information on what data and models or methods were used. This method would help down select which models or methods should be applied to the dataset that we were able to obtain.

4.1 Natural Language Processing

Throughout this research, Natural Language Processing (NLP) methods were utilized to further the work that could be done. As it has been mentioned, predicting CHF and impact of risk factors has been studied over the years and many diverse types of models have been applied to different data. NLP was used to help determine which models this research should continue to evaluate, and which ones have not been as commonly explored. There are three many aspects from the previous studies that were looked at to see what has been done and where can further investigation begin. The same method was used to pull all three pieces of information. This included data used,

and methods or models implemented in the research. The third piece of information utilized checks to see if there are factors that can be found across multiple research papers that have an impact on CHF. Model selection for the predictions done for this research was determined by which common models were looked at across the different studies and included a few methods that have been overlooked or not yet fully explored for this topic. Comparisons were drawn on the factors or impactors that were most common across the research papers and the results that were produced from the UCI dataset.

4.1.1 Tools and Python Libraries

For creating the prediction of CHF on the UCI dataset, we used Python version 3.5 with different python libraries like SpaCy, Natural Language Toolkit (NLTK) and PyPDF2 to ingest, clean, and preprocess article text from previous research to help accomplish NLP analysis in determining key words or phrases to improve research done in the future. PyPDF2 is a library that is a toolkit within Python to work with PDF files. Its capabilities include extracting document information, splitting or merging documents and many more applications to manipulate and utilize PDF files. This will be the primary method used in this research to be able to extract the information from the previous research papers compiled. Multiple libraries were used to help clean the text to make it ready for analysis. These libraries include NLTK, Regular Expression (RE), and Unicode Data. NLTK is a leading platform for building Python programs to work with human language data. There are a few libraries that were used to further prep the text and benefit the analysis. Being able to capture phrases that make more sense than the individual word is a key aspect in understanding themes or word occurrences utilizing a python library to create these phrases is important. There are a few diverse types of phraser methods common in python libraries but the one that was used for this analysis was the Gensim's Phraser. The Gensim's Phraser detects word phrases by utilizing a large corpus. The combined forces of the python libraries were utilized to perform named entity recognition or entity analysis and topic modeling. Entity analysis inspects the given text for known entities and returns information about those entities [26]. Additional topic modeling was performed and visualized to further understand the text within these research papers.

4.1.2 Application

Two pipelines were created to iterate through each of the research papers that were looked at for this research. The first pipeline that was created was to perform topic modeling with the Latent Dirichlet Algorithm. The second pipeline was created to look at a Name Entity Recognition analysis and visualize the most common dataset, models and factors found within the one-hundred research papers. Within these pipeline steps were taken to extract the information from the PDF files, clean and preprocess the raw text from those files, transforming the data, preform the different analyses, and visualize those outcomes. Although the pipelines have similar steps the packages used

for the individual analyses require different formatting, so it was simpler to have all work done for the sperate analyses.

4.1.2.1 Topic Modeling

For the topic modeling, the focus was solely on the text data from each research paper therefore the metadata was dropped. The first step of this pipeline was to extract the information from the PDF files and save them into a Pandas data frame. This was accomplished in two ways. The majority of the PDFs were able to have the information successfully extracted and made into raw text using the PYPDF2 library. The PDFFileReader function from this library was used to loop through each paper by traversing through each page of the paper and storing the data. The first page of each article was stored in one list while the rest of the page was stored in another list. These two lists where then combined to create the data frame. There were a few PDFs that failed to extract the information due to security settings within the properties of the PDF. Google Cloud Vision API (Application Programming Interface) was leveraged to use the OCR capability to retrieve the remaining PDF research paper text information.

The next and most crucial step of this pipeline was the cleaning and preprocessing of the raw text pulled from the PDF files. The steps that were taken in both analyses are common steps used when cleaning text data used for NLP concepts. These common tasks include tokenization, removals, and transformations of words. The first task of cleaning the text was to remove emails, new line characters, and single quotes by using the Regular Expression (re) function. This function allows the user to specify the pattern to be removed. After those items were removed, genism was used to tokenize the words within the sentences. Tokenization is the process in which the documents are split into words or small phrases that are called tokens. Tokens are a building block for processing and modeling of text. Tokenization was applied to each sentence within the papers by splitting the sentences into a list of words using gensim's simple_preprocess() function. The simple_preprocess function has a few built-in functions that were completed during the tokenization. These functions include transforming all uppercase to lowercase and removal of accented characters and punctuations by setting the deacc to true. The tokens were then converted into a list and additional preprocessing took place. This preprocessing included removal of stop words, creating the bigrams and trigrams, and lemmatization. Bigrams and trigrams, which are phrases with two or three words respectively, are created to make the clustering of the topics easier and better. There are many ways to create n-grams, but this analysis used the Gensim's Phrases model to create both the bigrams and trigrams models. There were two parameters indicated in the bigrams model with the min_count set to five and the threshold set to one hundred. For the trigram model, the bigram set was used with one parameter being selected, which was the threshold being set to one hundred. Prior to the bigram and trigram models being applied the stop words or frequently used words that do not add content or value were removed. For this analysis, the stopword function from the NLTK corpus library was used which has a pre-set list of words to use. After evaluating the

data, additional words were added to the list of stop words. This list included the following: ['from', 'subject', 're', 'edu', 'use', 'not', 'would', 'say', 'could', '_', 'be', 'know', 'good', 'go', 'get', 'do', 'done', 'try', 'many', 'some', 'nice', 'thank', 'think', 'see', 'rather', 'easy', 'easily', 'lot', 'lack', 'make', 'want', 'seem', 'run', 'need', 'even', 'right', 'line', 'even', 'also', 'may', 'take', 'come', 'pg_ijcsmc', 'kolkata', 'eswa', 'otice', 'earch', 'ijarp', 'hpps', 'graph_show', 'ijsr']. The last step taken to prep the data was the lemmatization process which takes the words down to their root form while using surrounding words and vocabular to properly convert the words to the root. The spaCy library was utilized to complete this task. The spaCy library was loaded to utilize the efficiency over accuracy with the en_core_web_sm command and disabling the ner and parser capabilities for the topic modeling analysis. During the lemmatization process within the spaCy lemma function, each word was changed to its root form, keeping only nouns, adjectives, verbs, and adverbs. These Parts-of-Speech (POS) tags were kept because they are the ones that contributed the most to the meaning of the sentences.

At this point the data set was ready to be transformed into a corpus and dictionary to feed into a LDA model used for topic modeling. LDA is a generative statistical model that helps pick up similarities across a collection of different data parts. A key assumption to understand LDA is that it assumes that the documents evaluated are represented by a fixed number of topics, and those topics are a distribution of words. The two main inputs to the LDA topic model are the dictionary(id2word) and the corpus. The dictionary was created by leveraging the dictionary function within the genism library which mapped the list of lemmatized bigrams and trigrams to an integer id. The doc2bow function converted the processed text data into bag-of-words (BoW) which is a list of the tokens and their count or term-frequency. With the corpus and dictionary created a baseline LDA model was run using the following parameters: number of topics set to ten, random state for the reproducibility to 100, update every set to the default of one, chunk size was reduced to ten, the number of passes through the corpus during training was set to ten, the alpha metric was set to the default of symmetric, the iterations through the corpus was set to 100 and the per_word_topics was set to true to compute a list of topics. After applying the LDA model, visualizations were created to highlight the topics relationships. The python library, pyLDAVis, was used to visualize to aspects of topic modeling. Looking at Figure 4, the left-hand side is a global view to help understand how prevalent each topic is to another and the right-hand side containing bar charts that represent the terms that are most useful in interpreting the topic currently selected on the global view. This analysis allowed us to narrow down which documents to focus on for our research i.e., determined more of what was done or to validate our results.

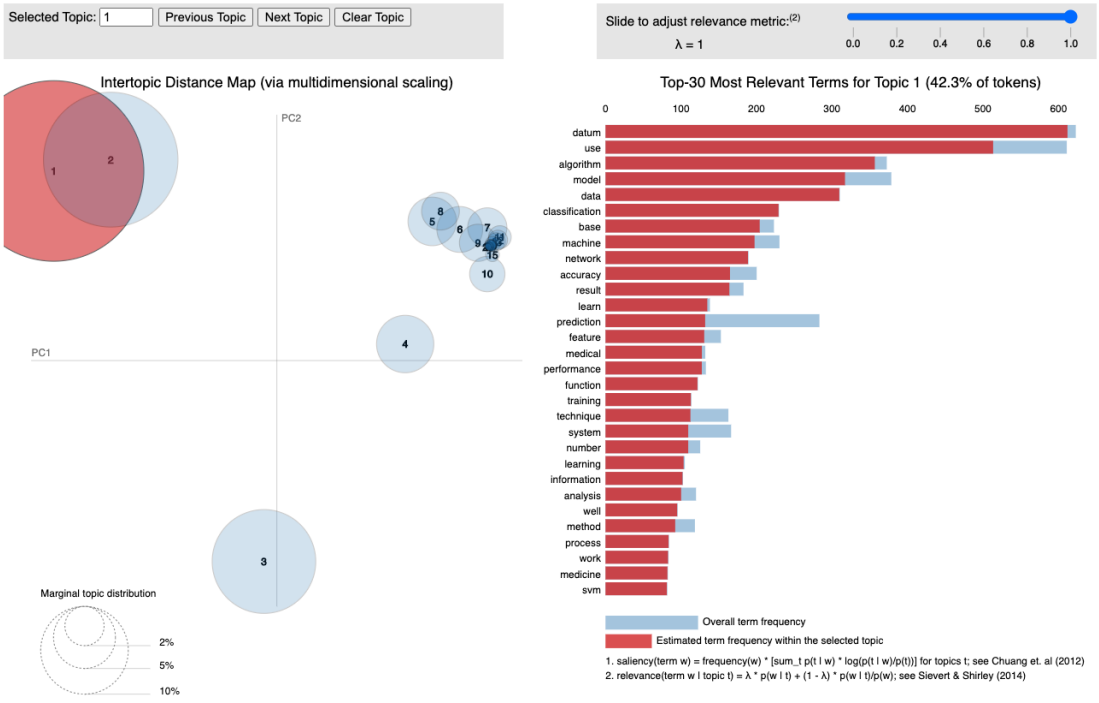

Figure 4: The layout of pyLDAvis, with the global topic view on the left, and the term bar charts with Topic 1 selected on the right. Linked selections allow users to reveal aspects of the topic-term relationships compactly.

4.1.2.2 Name Entity Recognition

The second analysis that was completed was to determine the distribution of certain words across the different research papers. In this piece of the research, we will be looking for familiar words that are grouped in certain datasets, models or methods and factors. This was accomplished by performing Named Entity Recognition (NER) and term frequency-inverse document frequency, also known as tf-idf. The python library spaCy, was used to assign labels and tags to groups of tokens. The spaCy function was loaded with the model (en_core_web_lg). Since the goal of this piece of analysis is to understand how often specific sets of words occur across the previous studies, tagging definitions were created. Three dictionaries were created one for each of the key items we were looking to extract and tag within the documents. Each dictionary has an overall word with a list of words that correspond to that word. The method or modeling tagger definition contains sixteen different kinds of methods that could be applied within machine learning research. For each of these methods the names of models, acronyms or common functions are listed as the tokens to look for. The other two dictionaries were set up in the same structure. The factor tagging dictionary contains fourteen factor

categories with lists of related words or acronyms. The third dictionary was created to capture the different types of data referenced in the papers. This dictionary contained nine dataset categories that have been collected for patients with the name or type of data that makes up that category. These dictionaries are an important piece to the analysis as the results are easily impacted based on whether certain words are in the dictionary. The tagger dictionaries used for this analysis can be found in the appendix.

The pipeline for this analysis has similar tasks as seen in the topic modeling pipeline but utilizes spaCy functions to complete the tasks. The first step in this pipeline was to extract the information from the PDF files and saved into a spaCy document object instead of a Pandas data frame. Custom boundaries were set to complete the sentence segmenter and tokenization during the creation of the spaCy document object. The spaCY pipeline was applied to the text data from the PDFs with all components activated and the custom boundaries that were set to create the spaCY document objects. The phrase matcher function within spaCy efficiently matched the dictionaries to words within the spaCy document object. This pipeline was applied three separate times one for the factors, methods and datasets.

4.2 Predicting Congestive Heart Failure

After understanding what common models were previously used, a pipeline was created for both the common and uncommon machine learning model to be evaluated on the UCI dataset. This original pipeline consists of twelve different kinds of models, some that were commonly used in previous studies such as Decision Tree Model and some uncommon methods when analyzing aspects of CHF. The set of twelve models include Decision Trees models, ensemble learning models, anomaly detection, Bayesian methods, Neural Networks, Regression analysis and dimensionality reduction methods. Based on the results of the first set of models, the models were down-selected further to optimize the top four performing models on the UCI dataset. The second pipeline included the top four performing models which were Linear Discriminant Analysis (LDA), Extra-Trees, Gradient Boost and Gaussian Naïve Bayes.

4.2.1 Tools and Python Libraries

For creating the prediction of CHF on the UCI dataset, Python version 3.5 was used to create the pipeline of models that were evaluated. Within Python, the Automated Tool for Optimized Modeling (ATOM) package was leveraged to quickly explore and experiment with different machine learning algorithms. ATOM contains a pipeline that easily handles data cleaning, feature selection, model selection and hyperparameter tuning and analyzing the results by producing easy to understand visuals. This allowed for quick analysis on the different models that were applied to the UCI dataset.

4.2.2 Models Evaluated

The models that were created in the pipeline span across diverse modeling methods and consisted of five ensemble learning methods, one decision trees, two anomaly detection methods, one Bayesian method, one dimensionality reduction method, one regression analysis, and one Artificial Neural Network.

The ensemble learning methods included the Ada Boost, Gradient Boosting Machine, XGBoost, Random Forest, and Extra-Trees. AdaBoost is a meta-estimator that begins by fitting a classifier/regressor on the original dataset and then fits additional copies of the algorithm on the same dataset but where the weights of instances are adjusted according to the error of the current prediction [27]. A Gradient Boosting Machine builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage n_classes_ regression trees are fit on the negative gradient of the binomial or multinomial deviance loss function. Binary classification is a special case where only a single regression tree is induced [28]. XGBoost is an optimized distributed gradient boosting model designed to be highly efficient, flexible and portable. XGBoost provides a parallel tree boosting that solve many data science problems in a fast and accurate way [29]. Random forests are an ensemble learning method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random forests correct for decision trees" habit of overfitting to their training set [30]. Extra-Trees use a meta estimator that fits several randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting [31]. Both the Random Forest and Extra-Trees algorithms are perturb-and-combine technique specifically designed for trees [32]. This means a diverse set of classifiers are created by introducing randomness in the classifier construction. The prediction of the ensemble is given as the averaged prediction of the individual classifiers.

The Decision Tree type of method was utilized in the pipeline which was the standard Decision Tree algorithm. Decision Tree is a single decision tree classifier/regressor [33].

The two anomaly detection methods included the K-Nearest Neighbors (KNN) and the Radius Nearest Neighbors (RNN). The Neighbors-based classification is a type of instance-based learning or non-generalizing learning: it does not attempt to construct a general internal model, but simply stores instances of the training data. The classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point [34]. KNN, as the name clearly indicates, implements the k-nearest neighbors vote [35]. Radius Nearest Neighbors implements the nearest neighbors vote, where the neighbors are selected from within a given radius. For regression, the target is predicted by local interpolation of the targets associated with the nearest neighbors in the training set [36].

The Bayesian method included the Gaussian Naïve Bayes (GNB) algorithm. The GNB is a supervised learning algorithm based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given

the value of the class variable [37]. GaussianNB implements the Gaussian Naive Bayes algorithm for classification. The likelihood of the features is assumed to be Gaussian.

The one dimensionality reduction method that was applied to the pipeline was the Linear Discriminant Analysis (LDA). Linear Discriminant Analysis is a classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes' rule. The model fits a Gaussian density to each class, if all classes share the same covariance matrix [38].

Only one regression analysis method was used within the pipeline. This method was the logistic regression algorithm. Despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum-entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function [39].

The final type of method that is found in the pipeline is one Artificial Neural Network called a Multi-layer Perceptron (MLP) classifier. MLP is a supervised learning algorithm that learns a function by training on a dataset and optimizes the log-loss function using LBFGS or stochastic gradient descent. Given a set of features and a target, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers [40]. There are two advantages when applying the MLP method as this can learn on non-linear models and in real-time using partial fit [41]. Although the real-time does not apply to the heart disease data this could be a benefit when looking at ECG or wearable devices that capture data for a patient that could be used in future studies.

4.2.3 Application

Since the ATOM package was used, majority of cleaning and preprocessing of the data is handled within the pipeline. There were a few steps that took place prior to pushing the data through the pipeline. The first step was to only keep the 14 variables required for this analysis. These 14 variables included are as follows: ['age', 'sex', 'chest_pain_type', 'resting_blood_pressure', 'cholesterol', 'fasting_blood_sugar', 'rest_ecg', 'max_heart_rate_achieved', 'exercise_induced_angina', 'st_depression', 'st_slope', 'num_major_vessels', 'thalassemia', 'condition']. The condition column was dropped as that was used as the predictor for the models. The remaining 13 variables were used to train the models and consisted of a combination of both categorical and numerical values. The eight categorical variables were manually mapped to numerical values starting at zero to n-1 where n is the number of values.

The data cleaning pipeline was created using the ATOMClassifier function. The 13 variables listed above were passed through the pipeline along with the predictor. The pipeline was initialized with the test size set to 30%, the random_state equal to 42 for reproducibility, the verbose set to two for all details to be printed and n_jobs set to –1 which allowed the pipeline to utilize all available CPUs.

The ways to complete data manipulation like imputation, scaling and encoding were set prior to the models being run. Although data imputation was not needed for this dataset, it was still added into the pipeline. The default methods were used from imputation which applied the k-nearest neighbor (KNN) algorithm to predict missing numeric values and the most common class to complete the categorical imputation. The scaling was left at the default setting of standard scaling. Leave on out encoding was utilized to encode the categorical variables which helped to level off an effect of outliers and created more diverse encoded values. Ordinal-encoding was applied to three features; sex, fasting_blood_sugar and exercise_induced_angina since these only contain two classes. The other five categorical variables utilized the one-hot encoding method. The last preprocessing step that needed to be called out was the feature selection which was applied by using the Principal Component Analysis (PCA).

Two pipelines for running models were created. The first was built for the twelve models and the f1 and recall were the metrics calculated but the models were optimized using the f1 score. The optimizer was initialized with 25 initial points and an additional 75 passed to find an optimal hyperparameters for each algorithm. The final parameter set for the first pipeline of models was the bootstrap aggregating or bagging. his technique creates several new data sets selecting random samples from the training set (with replacement) and evaluates them on the test set. For the first pipeline of models the bagging parameter was set to four. The second pipeline of models was created with top four performing models from the previous pipeline. Further optimization using the Bayesian optimizer was conducted in this pipeline again the optimizer was initialized with 25 initial points and had an additional 75 passed to find the optimal hyperparameters.

5 Results

5.1 NLP Results

The utilization of NLP concepts used on the previous studies gleaned a significant amount of insight on the use of data science and analytics techniques when looking at heart disease and in particular CHF. There seems to be limited improvements over the years on what types of data and methods are used. Through the topic modeling that was visualized in the Methods section above on NLP concepts, the bar charts on the right-hand side of Figure 4 showed that datum and data were in the top five words discussed or mentioned in the diverse groups of research papers.

5.1.1 Datasets

Based on the one-hundred research papers evaluated there was insight gleaned to help understand which data has been used and which has yet to be utilized when analyzing various aspects of CHF. From these results shown below in Figure 5, the most common data that has been used across numerous research papers and over time is the electronic health records. Around 80% of the documents evaluated utilized some electronic health records which includes the patient profiles and hospital or clinical electronic medical records. Both the UCI database and electro cardio dataset were referenced in around 30% of the previous studies used. The more specific data like wearable data, genetics, patient history and imaging all range from 5-15% discussion across the different studies. This result has shown that the analyses that have been conducted revolve around only a few kinds of data.

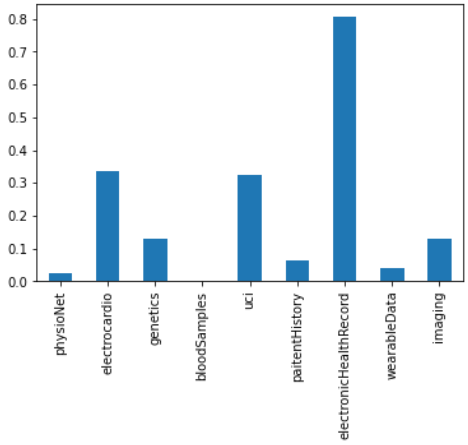

Figure 5: The distributed proportion of the type of data that were used in previous studies.

5.1.2 Models and Methods

The same analysis that was completed for the datasets was done to look at what models have been used in previous paper. The results shown in Figure 6 confirmed some of those assumptions while proved others to be inaccurate. Out of the one-hundred research papers none of those utilized statistical models like ANCOVA or Cox-Proportional Hazard Models. There were four types of models that were discussed in 40% or more of the research papers. The Decision Tree algorithms were the most common type of algorithm utilized when analyzing various aspects of CHF. A little over 60% of the research papers evaluated used some type of Decision Tree algorithms. Some type of linear classification was referenced the second most at about 55% of the papers. Newer machine learning algorithms like deep learning and Neural Networks are only referenced in around 1-10% of the one-hundred papers collected. There were quite a few types of models such as meta learning methods, unsupervised and semi

supervised learning, association rule learning, and hierarchical clustering that were not mentioned in any of the paper.

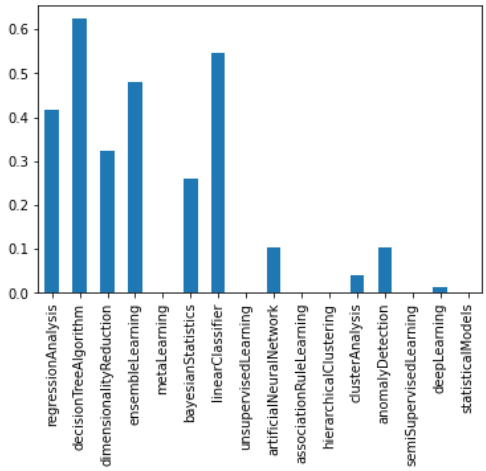

Figure 6: The distributed proportion of the types of methods used to during the analysis of different topics regarding Congestive Heart Failure.

5.1.3 Factors

The final piece of information that was extracted from the previous studies conducted was the factors that, according to the analyses, contributed to a patient's likelihood of experiencing CHF at some point in their life. Three of the top four factors cannot be contributed to the lifestyle of a patient. As you can see from Figure 7 below, age was the most common factor discussed across the different research papers. Around 70% of the documents mentioned age as a contributing factor to CHF. The other factors that shown up in around 40% of the documents as factors to CHF were diabetes, gender, and angina. The lifestyle or factors that a patient could showed up in about 5–25% of the documents that were evaluated. Factors like family history or genetics of a patient were only discussed as factors in about 5-10% of the documents that were pulled for this analysis. Factors that dealt with heart rhythms were not found in any of the documents collected. From this you can see that not a single factor is mentioned in all the studies, but there are some that show up a lot more often than others.

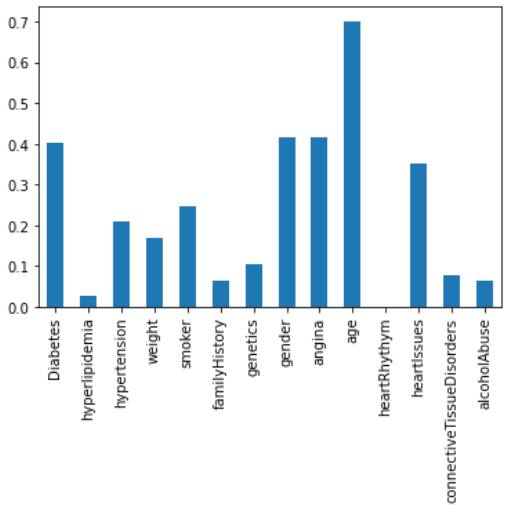

Figure 7: The distributed proportion of factors that lead to Congestive Heart Failure based on previous studies conducted.

5.2 Prediction Results

For comparing which machine learning models performed the best for each dataset we are focused on the precision and recall of the models, therefore F1 score will be one of the main performance metrics used to compare the models. Precision is the ratio between the True Positives and all the positives in the dataset. In other words, this would be the measure of patients that we correctly predicted to have CHF out of all the patients having CHF. The other metric that is important to consider is the Recall. The Recall represents the number of true positives divided by the number of true positives plus false negatives or, in simpler terms, the model's ability to find all the data points of interest in a dataset. The F1 score gives more weight to false negatives and false positives while not letting large numbers of true negatives influence the score. Since we are dealing with the healthcare industry, it is extremely important to focus on which metrics we are using to evaluate the best performing models. We want to avoid false negatives in this situation as this would lead to misdiagnosing a patient by saying they will not experience CHF and they end up having CHF. This could lead to a potential death that could have been avoided. NLP results narrowed the list of algorithms to include in the pipeline however linear classifiers and decision trees became the primary focus.

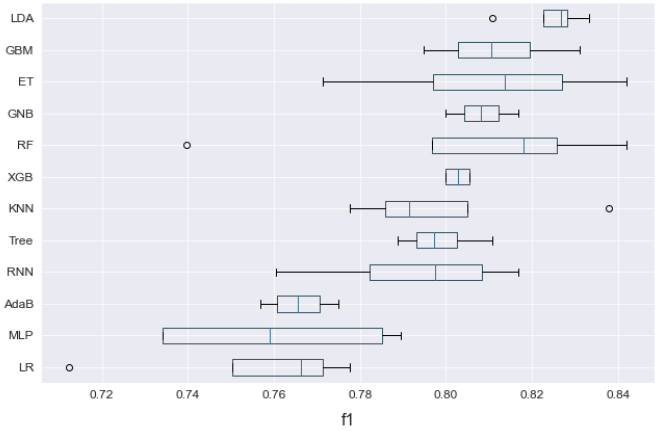

Figure 8: This shows the twelve models run in the first pipeline on the UCI dataset and the performance based on the f1 score.

In the baseline models that were built into the pipeline, there were six models that had a median f1 score of 0.8 or higher. All the median f1 scores for the twelve models seem to fall somewhere between 0.76 and 0.83. The LDA model performed the best out of all the models when looking at the f1 score. From Figure 8 above, there were some models that had a wide range of performances based on the parameters selected with outliers like the Extra-Trees, Random Forest, RNN and the MLP models. The worst performing models based on the f1 score were the liner regression and the MLP models. Another group of models that did poorly and had a wider range of performance was the two anomaly detection models, KNN and RNN. The standard decision tree fell in the bottom five models with a median f1 score of a little less than 0.8.

Since the ability to print the model details was turned on for the ATOM pipelines, we were able to see what the best parameters were set to for each model. The best parameter results can be seen in Figure 8 above by looking at the right most value for each of the models. All twelve models produced which parameters were set to something other than their default values and what that f1 score came out to be. The top four models that were chosen for further evaluation were the LDA, Gradient Boost Machine, Extra-Trees and Guassian Naïve Bayes. The dimensionality reduction LDA method had parameters set for the solver equal to least square regression (lsqr) and a shrinkage of 0.7. These parameters produced a f1 score of 0.8505 and a recall of 0.8389. The next set of methods are the ensemble learning algorithms which had more variability than LDA. The Gradient Boosting Machine was the top performer from a median f1 score standpoint in the ensemble learners, but the Extra-Trees algorithm best parameters outperformed the Gradient Boosting. The best parameters for the Gradient Boost Machine were learning rate set to 0.03, n_estimators set to 244, subsamle set to 0.6, criterion set to mse, min_samples_split set to 20, min_samples_leaf set to seven, max depth set to one, max features set to none, the ccp alpha set to 0.0 and the loss parameter set to deviance. These parameters produced an f1 score of 0.8419 and recall

of 0.8468 for the Gradient Boosting Machine. The Extra-Trees best f1 score of 0.8402 and recall of 0.8389 had the following parameters set: n_estimators: 498, criterion: entropy, max_depth: None, min_samples_split: 17, min_samples_leaf: 15, max_features: None, bootstrap: True, ccp_alpha: 0.009, 'max_samples': 0.8. The Guassian Naïve Bayes looked to have the least amount of variance in the f1 scores for the 100 iterations. The model with the default settings produced a f1 score of 0.7797 ± 0.0187 and a recall score of 0.7566 ± 0.0218.

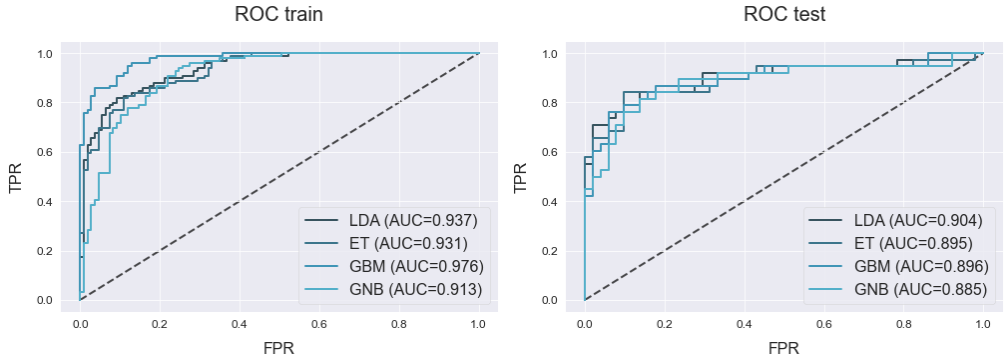

Figure 9: ROC curve that shows the performance of the models applied to the UCI data set at all the classification thresholds.

The ROC curve displays the trade-off between sensitivity and specificity for the ensemble models. Classifiers that give curves closer to the top-left corner indicate a better performance. The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test. The typical random classifier is expected to give points lying along the diagonal where the false positive rate equals the true positive rate. By looking at the ROC and AUC the LDA performed the best on the test data. From looking at Figure 9 above you can see that there was slight overfitting but for the most part these top four models performed well overall. Out of these four models it seems as though the Gradient Boosting Machine model may have experienced some overfitting. After reviewing both the ROC curve and f1 scores the LDA model seemed to have performed the best on the UCI dataset.

6 Discussion

As there is no cure, early detection for the likelihood of experiencing CHF and uncovering the level of impact a factor has can lead to improved preventative care or management plans to help reduce the severity of experiencing CHF. By focusing on predicting if a patient has a likelihood of experiencing CHF this leads to the potential of changing not only the way that physicians or clinics evaluate or manage patients with the CHF but what data they need to be collecting on patients. This study has the potential to impact clinical practices in the way that they look at treatment or management plans and data collection and distribution. This study is just the starting

point for understanding how to be able to answer the question of what the likelihood is that a patient will experience CHF later down the road based on all the different data points and what are ways to help prevent or mitigate this issue.

The most difficult aspect of this analysis is obtaining data that can be used as an all-encompassing look at an individual. There is extremely limited data that is publicly accessible regarding data that falls under the Health Insurance Portability and Accountability Act (HIPPA). For the data that you can find on public forums is either anonymized or not all collect for one given person. Even if you are within the health industry this data for one given individual may not be available or would need to go through a de-identification process. In addition, concerns regarding patient privacy and worry about revealing unfavorable institutional practices, hospitals and clinics have been extremely reluctant to allow access to clinical data for researchers from outside the associated institutions [20]. The issue of making certain data public is nothing new and has been a problem for a long time. There is a thought that data that is publicly funded like clinical data with individual patients has characteristics of a public good or public utility and that these data should be shared widely and used for the common good of improving the nation's health and healthcare system [42]. Health related data goes from a diverse group of both public and private systems that must weigh the cost and benefit of making these sources publicly available. There is benefit to multiple different groups in make increasingly large quantities of data available to the public which include patients, doctors, clinical trials, and informatics like graduate students or other professionals interested in healthcare. There are a few different reasons for why healthcare data as not become more widely accessible, but there is some ability to overcome those challenges. The main reasons for limited access to data is due to HIPAA with concerns of misuse of the data and more risk than benefit. A big piece for HIPAA and accessibility to the data revolves around the concept of de-identification. The current processes in place and the different improvements or methods that could be leveraged will be discussed as ways to increase the data accessibility problem.

The first challenge to overcome in attempting to make this data more easily accessible to the public is fully understanding HIPAA data and the laws and regulations to ensure that all parties are satisfied with the outcome. The data that is under HIPAA is considered protected health information that is individually identifiable information relating to the past, present or future status of the individual that was given to any provision of healthcare, payments, or healthcare services [46]. The healthcare industry is leap and bounds behind in exploring the full benefits of data analytics mainly due to concerns of HIPPA breaches. Benefits of utilizing data analytics have been acknowledge but there is still tension between the development of data analytics and privacy concerns under HIPAA. Although there have been improvements in the security of patient information, the concern is to keep this data HIPPA compliant. Under the Privacy Rule, health information and data that does not identify an individual or has been "de-identified" is no longer protected under HIPAA. The de-identification process is a way to keep healthcare data HIPPA compliant and allow for other to utilize this information. There are numerous methods in which de-identification can be accomplished from different methods or combinations of methods or tools that have implemented de-identification methods. There are two main ways that are provided by the Privacy Rule; the Expert Determination Method and the Safe Harbor method. The first method requires expert knowledge on both the information and statistics and

could led to each entity having very different de-identification processes. The second method is a more common approach with just the removal of specific data. The success of the de-identification process depends not only on the competency of the automatic de-identification systems, but also on the competency, dedication, and discipline of the institution that is responsible for de-identification of protected health information [43].

The difficulty of obtaining multiple different datasets led to being creative with how to approach the goal to answer a vital question of what data can help predict what a patient's likelihood of being diagnosed with CHF and can data on a patient help tailor a treatment or preventative plan to reduce the chance of progressing the disease or extending the life expectancy for an individual that is diagnosed with CHF. Although we were unable to answer our original goal of utilizing the full picture of a patient's health and not only predicting the likelihood that they could experience CHF but then give suggestions on management preventative plans was not met, we were able to gain insight into what has been done in a substantial amount of prior research conducted in a short amount of time. Besides gaining the insight into prior studies this research was able to highlight the gaps in data that have not been utilized and methods that could be investigated when analyzing CHF. Another immediate takeaway from this research is knowing what the common factors are that have led to CHF.

With conducting both the NLP analysis on the previous studies and the predictions using a diverse range of methods on a common data set used in heart related analyses we were able to create a starting point for other researchers to either utilize the NLP work to further their research or build upon what predictions were completed on a common data set and apply them to other data sets. The NLP analysis did highlight aspects of research that were surprising but overall was what we expected to see. When it came to the dataset analysis it was not a big surprise to see that Electronic Health Records were the most common dataset used as this type of data is extremely common information that is captured for all patients. Electronic health records include any data that is considered basic profile information of a patient or electronic medical record. This research was able to highlight that one of the most common datasets that has been used for analysis dealing with CHF is the UCI Cleveland dataset. This was slightly surprising with it being around one third of research, but this dataset is one of the only publicly available data more research papers will try to utilize this information. Utilizing wearable data and imaging is a new idea within the last decade so there might not be as many studies completed on these types of data. The electro cardio data is another set of data that seems to be used a decent amount of time for the research papers that were looked at. This could be slightly skewed based on the keywords used in the dictionary created for this dataset. A previous run resulted in having a large amount of blood sample datasets due to having blood as a keyword. The results of the methods used in previous studies was surprising to see that the diverse types of some machine learning methods and lack of utilization of more statistical methods or newer machine learning methods. The initial thought was that a large amount of the models would fall into the statistical methods as those are extremely common in the medical industry and that some of the newer machine learning and artificial intelligence methods would not been as commonly used. This could be due to the research papers that were collected focusing on machine learning techniques and casting a wider net on the focus of papers could show more statistical models. Since the research papers that were collected had a primary focus on which on machine learning methods for predicting CHF, it comes

as no surprise that most of the papers had success with a few different machine learning models. The top methods are utilized on the UCI dataset for predicting CHF. Although the research papers range from numerous years including the past few years, there are some newer machine learning methods that have not been utilized and some like deep learning have been used an exceedingly small amount. Even though trying to predict various aspects of CHF is nothing new the methods that have been used have not really changed over the years. The findings on the common factors across the studies did show that majority of the common factors are variables that cannot be controlled like age, gender and even some cases of diabetes. The initial thought was that there would be a larger number of factors like weight, diet, exercise, and other lifestyle factors contributing to experiencing CHF. This could be due to the type of data that was used for these analyses and could change based on the papers that are evaluated. For the prediction analysis, we were able to run multiple models with promising results on a small dataset. As the UCI dataset contributes to around 30% of the research papers should see some similarities between the top methods applied to CHF analyses and the most common factors. The top performing methods on the UCI included two ensemble learning methods, one Bayesian statistics method, and a dimensionality reduction method. The ensemble learning methods came as no surprise as being top performing models as they were commonly referenced in the previous studies, but the dimensionality reduction and Bayesian methods were only referenced in 25-30% of the studies. This percentage of references could be all studies that utilized the UCI datasets or other datasets.

There are types of improvements that could be done for this research. One of the ways is to leverage the work that has been done and modify different aspects of the analyses done. For the NLP analysis, fine tuning the dictionaries as to what is captured for each of the types of data, expansion of the methods used and including additional words related to those methods and building out the factors dictionary to include uncommon factors that could contribute to CHF. The expansion and fine tuning of these dictionaries can be accomplished by either using the topic modeling analysis and determining other common keywords or utilizing a subject matter expert within the field to ensure that the dictionaries are as accurate as possible to get the best insight. Another improvement that could be added to this research is searching for treatment or preventative plans that are common across the research completed. The other way that this research could be improved is by expanding the selection of papers collected as data. This was done manually by searching for analytics or machine learning regarding CHF but a creation to scrap all the papers from the different medical journal sites would increase the data size and give a better and more accurate depiction of what has been accomplished with analyzing CHF. Other improvements to the research and analysis can be accomplished by improving the access to redacted universal health care data. Obtaining data from regions where patients have access to universal health care would be ideal. A patient medical history overtime would provide better insight into the leading factors that contribute to CHF and how they manifest overtime. Prevention of CHF can start as early as childhood. Making this data universally available would allow a larger number of researchers and data scientist to bring a different view, methods, and analysis. Without access to additional data, it will be hard to improve on what has been published to date however the inclusion of imagery from body scans could be the change agent that is needed. Blood analysis is used to determine whether a patient has

cancer or not, but the severity of cancer cells can only be determined via a cat scan. Further research into predicting CHF should also include image processing and classification. The work that has been completed in this research can easily be utilized on many different aspects in the healthcare industry with minimal adjustments. This application of NLP creates a starting point for research on any analysis that could be looked at in the healthcare industry.

7 Conclusion

Although the initial goal of determining if a patient would experience CHF at some point in their life based on a variety of data was not accomplished due to limitations with access to data, this research is a starting point to understand what is needed to be able to accomplish that goal. With the sensitivity of healthcare data and the lack of public availability the healthcare industry is not able to utilize machine learning and data analytics to it full capability. An area of focus needs to be on implementing better or more standard ways to de-identify data and give access to the public so that they can help make tremendous contributions towards improving the healthcare industry. This research was able to highlight the lack of diverse data usage and model application regarding the analysis and prediction of CHF. This showed that although machine learning is being utilized in the healthcare industry little advancements in analysis for CHF have been made. Although the algorithms and datasets utilized have not seen many changes over the years these kinds of models showed promising results on the common UCI dataset. By working to create more publicly available data, the initial goal of decreasing the morbidity and mortality rates by predicting the likelihood of experiencing Congestive Heart Failure and ultimately specified preventative and management care plans would be within reach. Ultimately, this research can become the steppingstone for improvements to predicting Congestive Heart Failure and the utilization of machine learning across the healthcare industry with access to the right data.

Acknowledgments. Alexandra Norman. – Capstone Participant. James Harding. – Capstone Participant. Daria Zhukova. – Capstone Advisor. Jacquelyn Cheun. – Capstone Professor.

References

1. Cardiovascular diseases. (n.d.). Retrieved September 3, 2020, from https://www.who.int/health-topics/cardiovascular-diseases/
2. Habibović, M., Broers, E., Piera-Jimenez, J., Wetzels, M., Ayoola, I., Denollet, J., & Widdershoven, J. (2018, February 8). Enhancing Lifestyle Change in Cardiac Patients Through the Do CHANGE System ("Do Cardiac Health: Advanced New Generation Ecosystem"): Randomized Controlled Trial Protocol. Retrieved October 23, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5824100/
3. George, J., Rapsomaniki, E., Pujades-Rodriguez, M., Shah, A. D., Denaxas, S., Herrett, E., . . . Hemingway, H. (2015). How Does Cardiovascular Disease First Present in

Women and Men? *Circulation, 132*(14), 1320-1328. doi:10.1161/circulationaha.114.013797

4. Understanding Congestive Heart Failure. (n.d.). Retrieved September 15, 2020, from https://www.baylorhearthospital.com/Understanding-Congestive-Heart-Failure.html
5. Larrabee, B. (Ed.). (2019, January 03). Congestive heart failure basics. Retrieved October 20, 2020, from https://wa.kaiserpermanente.org/healthAndWellness?item=%2Fcommon%2FhealthAndWellness%2Fconditions%2FheartDisease%2FchfBasics.html
6. Brindles Lee Macon and Kristeen Cherney. (2018). Congestive heart failure: Types, causes, stages, and treatment. Healthline. https://www.healthline.com/health/congestive-heart-failure
7. Heart Failure: Risk Factors. (n.d.). Retrieved from https://www.universityhealth.org/heart-failure/risk-factors/
8. Betihavas, V., & Frost, S. A. (2015). Predicting Absolute Risk for Unplanned Cardiovascular Readmissions in Adults With Chronic Heart Failure. Journal of Cardiac Failure, 21(8). doi:10.1016/j.cardfail.2015.06.372
9. Awan, S. E., Sohel, F., Sanfilippo, F. M., Bennamoun, M., & Dwivedi, G. (2018). Machine learning in heart failure. *Current Opinion in Cardiology, 33*(2), 190-195. doi:10.1097/hco.0000000000000491
10. Jia, X., Baig, M. M., Mirza, F., & Gholamhosseini, H. (2019). A Cox-Based Risk Prediction Model for Early Detection of Cardiovascular Disease: Identification of Key Risk Factors for the Development of a 10-Year CVD Risk Prediction. *Advances in Preventive Medicine, 2019*, 1-11. doi:10.1155/2019/8392348
11. Folsom, A. R., Shah, A. M., Lutsey, P. L., Roetker, N. S., Alonso, A., Avery, C. L., Miedema, M. D., Konety, S., Chang, P. P., & Solomon, S. D. (2015). American Heart Association's life's simple 7: Avoiding heart failure and preserving cardiac structure and function. The American Journal of Medicine, 128(9), 970-976.e2. https://doi.org/10.1016/j.amjmed.2015.03.027
12. Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., & Schneeweiss, S. (2020). Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Network Open, 3(1). doi:10.1001/jamanetworkopen.2019.18962
13. Scikit Learn, Decision Trees, https://scikit-learn.org/stable/modules/tree.html, Accessed 15 September 2020.
14. Scikit Learn, Gradient Tree Boosting, https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting, Accessed 15 September 2020.
15. Scikit Learn, Forests of randomized trees, https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomized-trees, Accessed 15 September 2020.
16. Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making, 20(1). doi:10.1186/s12911-020-1023-5
17. Yuan, H., Fan, X., Jin, Y., He, J., Gui, Y., Song, L., . . . Chen, W. (2019). Development of heart failure risk prediction models based on a multi-marker approach using random forest algorithms. *Chinese Medical Journal, 132*(7), 819-826. doi:10.1097/cm9.0000000000000149
18. Ge, Y., & Wang, T. J. (2012). Identifying novel biomarkers for cardiovascular disease risk prediction. *Journal of Internal Medicine, 272*(5), 430-439. doi:10.1111/j.1365-2796.2012.02589.x
19. Velupillai, S., Suominen, H., Liakata, M., Roberts, A., Shah, A. D., Morley, K., . . . Dutta, R. (2018). Using clinical natural language processing for health outcomes

research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics, 88*, 11-19. doi:10.1016/j.jbi.2018.10.005

20. Chapman, W. W., Nadkarni, P. M., Hirschman, L., D'Avolio, L. W., Savova, G. K., & Uzuner, O. (2011). Overcoming barriers to nlp for clinical text: The role of shared tasks and the need for additional creative solutions. *Journal of the American Medical Informatics Association, 18*(5), 540-543. doi:10.1136/amiajnl-2011-000465
21. Torii, M., Fan, J., Yang, W., Lee, T., Wiley, M. T., Zisook, D. S., & Huang, Y. (2015). Risk factor detection for heart disease by applying text analytics in electronic medical records. *Journal of Biomedical Informatics, 58*. doi:10.1016/j.jbi.2015.08.011
22. Cho, H. (2019, June). Topic modeling. Retrieved February 31, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6590877/
23. UCI Machine Learning Repository, Heart Disease Data Set Cardiovascular Diseases, https://archive.ics.uci.edu/ml/datasets/heart+disease, Accessed 15 September 2020.
24. Mckee, P. A., Castelli, W. P., Mcnamara, P. M., & Kannel, W. B. (1971). The Natural History of Congestive Heart Failure: The Framingham Study. *New England Journal of Medicine, 285*(26), 1441-1446. doi:10.1056/nejm197112232852601
25. Scikit Learn, Permutation feature importance, https://scikit-learn.org/stable/modules/permutation_importance.html, Accessed 15 September 2020.
26. Natural language api basics | google cloud. (n.d.). Retrieved February 07, 2021, from https://cloud.google.com/natural-language/docs/basics#:~:text=Entity%20sentiment%20analysis%20inspects%20the,positive%2C%20negative%2C%20or%20neutral.
27. Tvdboom. (n.d.). AdaBoost (ADAB). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/adab/
28. Tvdboom. (n.d.). Gradient boosting Machine (GBM). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/gbm/
29. Tvdboom. (n.d.). XGBoost (xgb). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/xgb/
30. Tvdboom. (n.d.). Random forest (rf). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/rf/
31. Tvdboom. (n.d.). Extra-trees (et). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/et/
32. Scikit Learn, AdaBoost, https://scikit-learn.org/stable/modules/ensemble.html#adaboost, Accessed 15 September 2020.
33. Tvdboom. (n.d.). Decision tree (tree). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/tree/
34. Scikit Learn,Nearest Neighbors Classification, https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification, Accessed 15 September 2020.
35. Tvdboom. (n.d.). K-Nearest neighbors (KNN). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/knn
36. Tvdboom. (n.d.). Radius nearest Neighbors (RNN). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/rnn/
37. Tvdboom. (n.d.). Gaussian naive bayes (GNB). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/gnb
38. Tvdboom. (n.d.). Linear discriminant analysis (lda). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/lda/
39. Tvdboom. (n.d.). Logistic regression (lr). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/lr/
40. Tvdboom. (n.d.). Multi-layer perceptron (MLP). Retrieved February 10, 2021, from https://tvdboom.github.io/ATOM/API/models/mlp/

41. Scikit Learn, Neural network models (supervised), https://scikit-learn.org/stable/modules/neural_networks_supervised.html, Accessed 15 September 2020.
42. Institute of Medicine (US) Roundtable on Value & Science-Driven Health Care. (1970, January 01). Healthcare data: Public good or private property? Retrieved March 07, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK54304/
43. What is considered protected health information under HIPAA? (2020, August 18). Retrieved March 02, 2021, from https://www.hipaajournal.com/what-is-considered-protected-health-information-under-hipaa/#:~:text=Under%20HIPAA%2C%20protected%20health%20information,provision%20of%20healthcare%2C%20payment%20for

Appendix:

Dictionaries for Document Tagging

```
## Document Methods Tagging Definitions
regressionAnalysis = ["logistic regression", "ordinary least squares regression", "olsr", "linear regression", "stepwise regression", "multivariate adaptive regression splines", "mars", "regularization algorithm", "ridge regression", "least absolute shrinkage and selection operator", "lasso", "elastic net", "least-angle regression", "lars", "probabilistic classifier", "naive bayes classifier", "binary classifier", "linear classifier", "hierarchical classifier"]

decisionTreeAlgorithm = ["decision tree", "classification and regression tree", "cart", "iterative dichotomiser 3", "id3", "c4.5 algorithm", "c5.0 algorithm", "chi-squared automatic interaction detection", "chaid", "decision stump", "conditional decision tree", "id3 algorithm", "random forest", "sliq"]

dimensionalityReduction = ["canonical correlation analysis","cca","factor analysis","feature extraction", "feature selection",ion", "feature selection", "independent component analysis", "ica", "Linear discriminant analysis"," (LDA)", "multidimensional scaling","mds", "Non-negative matrix factorization","nmf", "partial least squares regression","plsr)", "principal component analysis","pca", "principal component regression", "pcf", "projection pursuit", "sammon mapping", "t-distributed stochastic neighbor embedding","(t-SNE)"]bor embedding"," (t-SNE)"]

ensembleLearning = ["adaboost", "boosting", "bootstrap aggregating","bagging", "ensemble averaging", "gradient boosted decision tree","gbdt" "gradient boosting machine","gbm", "random forest", "stacked generalization","blending"]

metaLearning =["inductive bias", "metadata"]

bayesianStatistics = ["bayesian knowledge base", "naive bayes", "gaussian Naive bayes", "multinomial naive bayes", "averaged one-dependence estimators","aode", "bayesian belief network","bbn", "bayesian network","bn"]

linearClassifier = ["fishers linear discriminant", "linear regression", "logistic regression", "multinomial logistic regression", "naive bayes classifier", "perceptron", "support vector machine", "svm"]
```

```
  unsupervisedLearning = ["expectation-maximization algorithm", "vector
quantization", "generative topographic map", "information bottleneck method"]

  artificialNeuralNetwork = [ "feedforward neural network", "extreme learning
machine", "convolutional neural network", "recurrent neural network", "long
short-term memory", "lstm", "logic learning machine", "self-organizing map"]

  associationRuleLearning = [ "apriori algorithm", "eclat algorithm", "fp-
growth algorithm"]

  hierarchicalClustering = [ "single-linkage clustering", "conceptual
clustering"]

  clusterAnalysis = [ "birch", "dbscan", "expectation-maximization", "em",
"fuzzy clustering", "hierarchical clustering", "k-means clustering", "k-
medians", "mean-shift", "optics algorithm"]

  anomalyDetection = ["k-nearest neighbors classification", "k-nn", "local
outlier factor"]

  semiSupervisedLearning = ["active learning", "generative models", "low-
density separation", "graph-based methods", "co-training", "transduction"]

  deepLearning = ["deep belief networks", "deep boltzmann machines", "deep
convolutional neural networks", "deep recurrent neural networks", "hierarchical
temporal memory", "generative adversarial networks", "deep boltzmann machine",
"dbm", "stacked auto-encoders"]

  ## Data Tagging Definition
  physioNet = ["physionet", "mit database", "mit", "bih", "bih
database"]

  electrocardio = ["electrocardiogram", "electrocardiograph", "ekg",
"ecg", "rr intervals", "electrical activity", "heart rate"]

  genetics = ["genetics", "gene", "biomarker"]

  bloodSamples = ["bloodwork", "blood samples"]

  uci = ["uci", "cleveland data"]

  paitentHistory = ["paitent history", "family history", "family
medical history", "paitent medical history", "past diagnosis"]

  electronicHealthRecord = ["electronic health record", "basic profile
information", "profile information", "ehr", "emr", "diagnosis",
"electronic medical record"]

  wearableData = ["apple watch", "heart monitor", "wearable"]

  imaging = ["heart images", "imaging"]

  ## Factor Tagging Definition
  Diabetes = ["diabetes", "basal secretion", "blood glucose level",
"bolus secretion", "diabetes mellitus", "dka", "diabetic ketoacidosis",
```

```
"glucose intolerance", "a1c", "glucose", "hyperglycemia", "diabetic",
"hypoglycemia", "blood glucose", "pre-diabetes"]

  hyperlipidemia = ["high ldl",  "high chol", "high cholesterol", "bad
cholesterol", "lipid disorder", "hypercholesterolemia", "high density
lipoprotein", "hdl cholesterol"]

  hypertension = ["high bp","high blood pressure", "htn", "ht"]

  weight = ["obesse", "bmi", "overweight", "obesity"]

  smoker =["smoker", "smoking", "tobacco"]

  familyHistory = ["family history"]

  genetics = ["genes", "genetics", "biomarkers"]

  gender = ["male", "female", "gender"]

  angina = ["angina", "chest pain"]

  age = ["age"]

  heartRhythm = ["abnormal EKG", "abnormal ECG", "EKG abnormality"]

  heartIssues = ["cad", "coronary artery disease", "previous heart
attack", "heart valve disorder"]

  connectiveTissueDisorders = ["connective tissue disorder", "systemic
lupus erythematosus", "sarcoidosis", "amyloidosis"]

  alcoholAbuse = ["alcohol abuse", "alcohol consumption", "drug abuse"]
```

Articles

- Congestive heart failure, Towards a comprehensive treatment, S. H. Taylor
- Pathoohvsioloev and Current Therapy of Congestive Heart Failure, William W. Parmley, MD, FACC
- Advanced heart failure: a position statement of the Heart Failure, Association of the European Society of Cardiology
- Diagnosis and Evaluation of Heart Failure, Michael King, MD; Joe Kingery, Do; and Baretta Casey, MD, MPH
- Congestive Heart Failure, Athena Poppas and Sharon Rounds
- PBM-MAP Clinical Practice Guideline for the Pharmacologic Management of Chronic Heart Failure in Primary Care Practice, Pharmacy Benefits Management Strategic Healthcare Group and Medical Advisory Panel
- Chronic heart failure in adults: diagnosis and management, NICE guideline
- Investigation and management of congestive heart failure, Bruce Arroll, Robert Doughty, Victoria Andersen
- Chronic heart failure, Australian Prescriber, Volume 40: Number 4: August2017
- A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction, Dr. Durairaj.M, Sivagowry.S

- Improvising Heart Attack Prediction System using Feature Selection and Data Mining Methods, B.Kavitha
- Comparing Performance of Data Mining Algorithms in Prediction Heart Diseases, Moloud Abdar
- Machine learning prediction in cardiovascular diseases: a meta-analysis, Chayakrit Krittanawong
- Heart Disease Prediction using Machine Learning Techniques, Devansh Shah · Samir Patel · Santosh Kumar Bharti
- Heart Diseases Prediction using Deep Learning Neural Network Model, Sumit Sharma, Mahesh Parmar
- Artificial Intelligence, Machine Learning, and Cardiovascular Disease, Pankaj Mathur
- Accurate Prediction of Coronary Heart Disease for Patients With
- Hypertension from Electronic Health Records with Big Data and Machine-Learning Methods:Model Development and Performance Evaluation, Zhenzhen Du
- Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward an Open-Source Solution Aravind Akella, Vibhor Kaushik, Qualicel Global Inc., Huntington Station, NY 11746
- Implementation of Machine Learning Model to Predict Heart Failure Disease, Fahd Saleh Alotaibi
- Heart disease prediction using machine learning techniques: A survey Article in International Journal of Engineering & Technology · March 2018
- Heart Disease Prediction using Machine Learning, Apurb Rajdhan
- Heart Disease Prediction using Machine Learning S.Nandhini
- Heart disease prediction using machine learning algorithms, Harshit Jindal et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1022 012072
- Heart Disease Prediction Using Machine learning and Data Mining Technique, Jaymin Patel
- Analysis of Weka Data Mining Techniques for Heart Disease Prediction System, Basma Saleh
- Heart Disease Prediction Using Machine Learning Techniques, Galla Siva Sai Bindhika
- Effective Heart Disease Prediction using Distinct Machine Learning Techniques, N. Suganthi
- Heart Disease Prediction Using Effective Machine Learning Techniques, Avinash Golande
- Heart Disease Prediction using Machine Learning Models, Ruban, Vivek, Krithi
- Early and accurate detection and diagnosis of heart disease using intelligent computational Model, Yar Muhammad
- Predicting the presence of heart disease using Machine Learning, Akshay Jayraj Suvarna
- A Review on Heart Disease Prediction using Machine Learning, Akhand Pratap Singh
- Prediction of Heart Disease Using Machine Learning Algorithms, Sonam Nikha
- Heart Disease Predictin Using Machine Learning Techniques, Dr. S.V. Kogilvani
- Heart disease prediction using machine learning techniques: a survey, V.V. Ramalingam
- Heart Disease Prediction with Machine Learning Approaches, Megha Kamboj
- Heart Disease Prediction using Machine Learning, Riddhi Kasabe
- An Efficient Convolutional Neural Network for Coronary Heart Disease Prediction, Aniruddha Dutta
- Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review, Animesh Hazra
- HPPS: Heart Problem Predicting System Using Machine Learing, Nimai Chand, Das Adhikari
- Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, Youness Khourdifi
- Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques, R. Chitra
- Hybrid Machine Learning Techniques for Heart Disease Prediction, S. Sharanyaa
- Heart Disease Prediction System Using Supervised Learning Classifier, R. Chitra
- A Comprehensive Review of Heart Disease Prediction Using Machine Learning, Dr Dilbag Singh

- Review of Heart Disease Prediction using Supervise and Unsupervised Machine Learning Technique, Pooja Gupta
- A Review on Heart Disease Prediction Using Machine Learning Techniques, Adil Hussain She
- Heart Disease Prediction using Logistic Regression Algorithm using Machine Learning, Reddy Prasad, Pidaparthi Anjali, S. Adil, N. Deepa
- Multiple Disease Prediction Using Different Machine Learning Algorithms Comparatively, Rudra A. Godse, Smita S. Gunjal, Karan A. Jagtap, Neha S. Mahamuni Prof. Suchita Wankhade
- Heart Disease Prediction Using Machine Learning Algorithm, Praveen Kumar Reddy M, T Sunil Kumar Reddy, S. Balakrishnan, Syed Muzamil Basha, Ravi Kumar Poluru
- Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab, A. S. Thanuja Nishadi
- Designing a Disease Prediction Model using Machine Learning, Ms.Jyoti Chandrashekhar Bambal
- Prediction System for Heart Disease Based on Ensemble Classifiers, Joshua Emakhu and Sujeet Shrestha
- Heart Disease Prediction Using Machine Learning, Siddhesh Iyer
- Heart Disease Prediction System Using Data Mining Techniques, Abhishe k Taneja
- Prediction of heart disease by classifying with feature selection and machine learning methods, Cengiz Gazeloğlu
- Heart Failure Prediction Using Machine Learning Techniques, Prasanta Kumar Sahoo
- Diagnosis of Heart Disease Using Data Mining Algorithm, Deepali Chandna
- Effectiveness of LSTMS in Predicting Congestive Heart Failure Onset, Sunil Mallya
- Clinical practice update on heart failure 2019: pharmacotherapy, procedures, devices and patient management. An expert consensus meeting report of The Heart Failure Association of the European Society of Cardiology, Petar M. Seferovic
- Living with heart failure, Resources to help you manage your heart failure, Cardiac Services BC
- The Future of Heart Failure Diagnosis, Therapy, and Management, James E. Udelson
- The causes, consequences, and treatment of left or right heart failure, Pablo Pazos-López, Jesús Peteiro-Vázquez, Ana Carcía-Campos, Lourdes García-Bueno, Juan Pablo Abugattas de Torres, Alfonso Castro-Beiras
- Novel biomarkers for cardiovascular risk prediction, Juan Wang
- Cardiorespiratory Fitness, Body Mass Index, and Heart Failure Mortality in Men Cooper Center Longitudinal Study, Stephen W. Farrell
- Risk Factors for Heart Failure: A Population-Based Case-Control Study, Shannon M. Dunlay, M.D.
- The Heart Failure Manual, Information on managing and improving your condition, Seton Health Care
- Heart Failure - Systolic Dysfunction, William E Chavey, MD
- Congestive Heart Failure: Provider's guide to diagnose and code CHF, Cigna HealthSpring
- Pathophysiology of Heart Failure Mathew Maurer, MD, Assistant Professor of Clinical Medicine Columbia University
- What to Expect: Living with Heart Failure, UPMC Hearth and Vascular Institute
- Heart failure Guidelines A concise summary for the GP, John J. Atherton PhD, MB BS, FRACP, FCSANZ, FESC
- Vericiguat in Patients with Heart Failure and Reduced Ejection Fraction, Paul W. Armstrong, M.D.
- Heart Failure, Article in South Dakota journal of medicine · October 2015
- Management of chronic heart failure, Healthcare Improvement Scotland, March 2016
- A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms, Amin Ul Haq

- Improving risk prediction in heart failure using machine learning, Eric D. Adler1
- Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes, Rishi J. Desai, MS, PhD
- Machine learning versus conventional clinical methods in guiding management of heart failure patients—a systematic Review, George Bazoukis
- Predicting survival in heart failure: a risk score based on 39,372 patients from 30 studies, Stuart J. Pocock1
- Predicting Death Due to Progressive Heart Failure in Patients With Mild-to-Moderate Chronic Heart Failure, Mark T. Kearney, DM
- Predicting Mortality Among Patients Hospitalized for Heart Failure Derivation and Validation of a Clinical Model, Douglas S. Lee, MD
- Predicting mortality in patients with heart failure: a pragmatic approach, M L Bouvy, E R Heerdink, HGM Leufkens, A W Hoes
- Clinical prediction of incident heart failure risk: a systematic review and meta-analysis, Hong Yang, Kazuaki Negishi, Petr Otahal, Thomas H Marwick
- Predictors of Congestive Heart Failure in the Elderly: The Cardiovascular Health Study, John S. Gottdiener, MD, FACC
- Predictors of mortality and morbidity in patients with chronic heart failure, Stuart J. Pocock
- Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques, Evanthia E. Tripoliti
- Detecting Congestive Heart Failure by Extracting Multimodal Features and Employing Machine Learning Techniques, Lal Hussain
- Machine learning prediction in cardiovascular diseases: a meta-analysis, Chayakrit Krittanawong
- Artificial intelligence for the diagnosis of heart failure, Dong-Ju Choi
- Evaluating risk prediction models for adults with heart failure: A systematic literature review, Gian Luca Di Tanna, Heidi Wirtz, Karen L. Burrows, Gary Globe
- Electron-Beam Tomography Coronary Artery Calcium and Cardiac Events, A 37-Month Follow-Up of 5635 Initially Asymptomatic Low- to Intermediate-Risk Adults, George T. Kondos
- The Future of Heart Failure Diagnosis, Therapy, and Management, James E. Udelson
- Risk Factors for Heart Failure: A Population-Based Case-Control Study, Shannon M. Dunlay
- Risk factors for and management of heart failure Thomas F. Luscher, MD, FESC
- Preventable causative factors leading to hospital admission with decompensated heart failure, A Michalsen, G König, W Thimme
- Factors Contributing to the Hospitalization of Patients with Congestive Heart Failure, Marshall H. Chin, MD, MPH, and Lee Goldman, MD, MPH
- Cardiorespiratory Fitness, Body Mass Index, and Heart Failure Mortality in Men Cooper Center Longitudinal Study, Stephen W. Farrell, PhD
- American Heart Association's Life's Simple 7: Avoiding Heart Failure and Preserving Cardiac Structure and Function, Aaron R. Folsom, MD