**A Case Study of Natural Language Processing**

**Applied to Clinical Notes for Detection of High Mortality Conditions**

Korey E. Bernhardt

Department of Data Science, University of Wisconsin-Stevens Point

DS 785: Capstone

Professor Alexander Korogodsky

December 11, 2022

**Table of Contents**

List of Tables 4

List of Figures 5

Abstract 6

Chapter 1: Introduction 7

Problem Statement 9

Background 9

Objectives and Purpose 10

Research Questions 11

Significance 11

Assumptions and Limitations 12

Chapter 2: Literature Review 12

NLP Overview and History 13

NLP Application in Healthcare 14

Recent NLP Healthcare Studies 15

Current Limitations and Challenges 16

Summary 17

Chapter 3: Methodology 17

Dataset Overview 18

Data Cleansing 19

General Data Cleansing 20

NLP Specific Data Cleansing 20

NLP Tools Overview 23

Scikit-learn 23

ScispaCy and UMLS 24

Disease Labelling 24

Classification Modeling Approach 27

Logistic Regression 28

Naïve Bayes 29

Decision Trees 29

Neural Networks 30

Model Testing and Verification Approach 31

Parameter Tuning 32

Classification Metrics 33

Summary 34

Chapter 4: Results and Findings 34

Descriptive Statistics 35

Supervised Learning Results 37

Unsupervised Learning Results 41

Comparison of Results 44

Summary 45

Chapter 5: Discussion 45

Extracting High Mortality Conditions from Clinical Notes 46

Unsupervised Learning Models for NLP of Clinical Notes 46

Scalability of NLP for Clinical Notes 47

Case Study Limitations 47

Recommendations for Future Research 48

Conclusion 48

References 50

Appendix A 55

Appendix B 57

Multiclass Classification Results 57

Binary Classification Results 61

Appendix C 65

# List of Tables

**Table 1** Medical Specialty Distribution 26

**Table 2** Hyperparameter Tuning Summary 33

**Table 3** Multiclass Logistic Regression Classification Report 39

**Table 4** Binary Logistic Regression Classification Report 40

**Table 5** Multiclass Neural Networks Training Classification Report 42

**Table 6** All Model Results 44

**Table A1** Aggregated Disease Mapping 55

[**Table B1** Multiclass Logistic Regression Classification Report 57](#_Toc121687113)

[**Table B2** Multiclass Naïve Bayes Classification Report 58](#_Toc121687114)

[**Table B3** Multiclass Decision Tree Classification Report 58](#_Toc121687115)

[**Table B4**  Multiclass Neural Networks Training Classification Report 59](#_Toc121687116)

[**Table B5** Multiclass Neural Networks Testing Classification Report 60](#_Toc121687117)

[**Table B6** Binary Logistic Regression Classification Report 61](#_Toc121687118)

[**Table B7** Binary Naive Bayes Classification Report 62](#_Toc121687119)

[**Table B8** Binary Decision Tree Classification Report 63](#_Toc121687120)

[**Table B9** Binary Neural Networks Training Classification Report 63](#_Toc121687121)

[**Table B10** Binary Neural Networks Testing Classification Report 64](#_Toc121687122)

# 

# List of Figures

**Figure 1** Count of Keywords per Transcription 19

**Figure 2** Feature Name Word Cloud 22

**Figure 3** *High Mortality Disease Label Distribution* 27

**Figure 4** *Cancer/Non-Cancer Label Distribution* 27

**Figure 5** Neural Network Example 31

**Figure 6** Median Transcription Word Count by Disease 35

**Figure 7** Word Count in Transcriptions 36

**Figure 8** Top 20 Feature Names by Disease 37

**Figure 9** Supervised Learning Model Metrics 38

**Figure 10** Multiclass Logistic Regression Confusion Matrices 40

**Figure 11** Binary Logistic Regression Confusion Matrices 41

**Figure 12** Unsupervised Compared to Supervised Learning Model Metrics 43

**Figure 13** ROC Curve and AUC Scores for Binary Models 45

**Figure B1** Multiclass Logistic Regression Confusion Matrices 57

**Figure B2** Multiclass Naïve Bayes Confusion Matrices 58

**Figure B3** Multiclass Decision Tree Confusion Matrices 59

**Figure B4** Multiclass Neural Networks Training Confusion Matrices 60

**Figure B5** Multiclass Neural Networks Testing Confusion Matrices 61

**Figure B6** Binary Logistic Regression Confusion Matrices 62

**Figure B7** Binary Naïve Bayes Regression Confusion Matrices 62

**Figure B8** Binary Decision Trees Confusion Matrices 63

**Figure B9** Binary Neural Networks Training Confusion Matrices 64

**Figure B10** Binary Neural Networks Testing Confusion Matrices 64

# Abstract

This project was an exploratory case study that took an in-depth look at using natural language processing (NLP) approaches to categorize unstructured data in the medical field, notably clinical notes. A look at the history of using NLP for clinical notes was completed, in addition to creating models to apply NLP to a dataset for further analysis. The primary question to be answered was whether high mortality conditions could be extrapolated from clinical notes. Additional research questions included whether unsupervised learning models could be used and whether the approach could be scaled to other diseases. Detecting this information can significantly benefit primary care providers, medical specialists, and ultimately patients as people rely more and more on multiple care providers to support their medical needs. Four models were used to compare results, including three supervised models and one unsupervised model. Multiclass and binary classification approaches were analyzed. While a binary logistic regression model performed the best, with 91% balanced accuracy and 92% weighted recall, an unsupervised neural networks model also achieved good results, with 89% balanced accuracy and 90% weighted recall. These results indicate that high mortality conditions can be extrapolated from clinical notes with a high degree of accuracy using unsupervised learning. The results can be scaled to additional diseases with further research.

*Keywords:* Natural language processing, clinical notes, high mortality disease classification, unsupervised learning, machine learning

# Chapter 1: Introduction

Missed and delayed medical diagnoses in the US are costly in multiple ways. Various sources estimate the financial costs are between $2 billion and $185 billion per year. The estimated cost of malpractice claims is $2 billion (How Misdiagnosis Is Increasing Healthcare Costs and How MORE Health Is Fixing the Problem, n.d.), while the estimated cost of inefficient services and missed prevention services is $185 billion (Medicine, 2012), per year. The human cost is estimated to be between 40,000 and 80,000 deaths per year caused by medical misdiagnoses (PinnacleCare, 2016). The cause can be attributed to many things, including physicians not having enough time per patient and miscommunication from specialists to primary care providers (Savage Law Firm, 2014).

Meanwhile, the volume of healthcare data that physicians can use to aid in diagnoses, reduce time to diagnose, and improve communication between physicians continues to rise. In 2018, healthcare data accounted for approximately 30% of all data worldwide (RBC Capital Markets | Navigating the Changing Face of Healthcare Episode, n.d.). It is also estimated that 80% of healthcare data is unstructured (Solutions, 2022). Unstructured healthcare data includes medical images, voice recordings, and medical notes. With the volume of data and benefits that can be realized, it is imperative to have a way to analyze unstructured medical data quickly and accurately.

One example of unstructured healthcare data is clinical notes. A physician typically writes or dictates clinical notes to capture details of an interaction with a patient. They document a visit, including findings, observations, diagnoses, and other information the clinician may find relevant about the visit. The notes are intended to be shared with other physicians, and some standardized formats are recommended to help clinicians understand them better. Even with standardization, it can be time-intensive for other clinicians to read and interpret a patient's clinical notes history to recognize potential health issues. Natural Language Processing (NLP) provides the bridge from unstructured clinical notes potentially hiding important information to structured, understandable translations of the contents of the clinical notes.

Simplistically, NLP is the ability of a machine to understand human language. This may be in the form of text or speech. While individual words may be easy to understand, language has deeper meanings, and understanding intent and emotions often play a role. Consider a word as basic as "right". On its surface, this word could mean correct, or it could mean a direction. Add intent and emotion, and now there are even more nuances. Is it a positive exclamation of shared understanding? Is it a sarcastic expression? Is it a speech pattern that's just keeping the conversation moving along? Add in different languages, dialects, slang, abbreviations, and misspellings, and this quickly becomes a highly complex undertaking.

While there are many applications for NLP, one way to use NLP is to label or classify data. In the context of medical data such as clinical notes, the classification may be based on domain, subdomain, diagnoses, medical conditions, and comorbidity identification, among others. This paper focuses on medical conditions, specifically a set of high mortality conditions. Classifying this type of information allows physicians to quickly identify areas of concern for a patient and can be cross-referenced with available structured data in a patient's electronic medical record (EMR) to determine specific patient risk. As patients rely more and more on multiple care providers to support their medical needs, quick detection of medical conditions from clinical notes can reduce the time spent by primary care providers and medical specialists. It also leads to more accurately detecting potential health concerns to improve patient outcomes.

## Problem Statement

Due to their unstructured nature, clinical notes capture information about a patient encounter that may be overlooked in favor of structured data that is easier to access and work with, such as ICD Codes that are intended to identify disease information for billing purposes. However, there can be gaps or inaccuracies in structured data that can lead to misdiagnosing a patient. A 2021 study comparing the identification of aortic stenosis using ICD codes to an NLP model evaluating echocardiogram reports demonstrated significantly increased accuracy of disease identification in the NLP model (Solomon et al., 2021). Using ICD codes, 59% of actual aortic stenosis cases were identified. Using NLP, 99% of actual cases were identified. The results of this study highlight the problem that reliance on structured data alone can lead to adverse patient outcomes.

Additionally, it is common for NLP applications to rely on supervised learning models where the development of the model relies on the clinical notes used for training the model being labeled with the correct target class. Labeling the notes can be a time-consuming process. The ability to use unsupervised learning to eliminate the need for pre-labeling is another area this paper analyzes in order to overcome this problem.

Lastly, many NLP projects for medical use are created for highly specialized research. The example cited concerning aortic stenosis is a prime example. While this type of research is also important, a more generalized approach that can be used as a framework for ongoing expansion is also needed.

## Background

Natural Language Processing is a relatively new data science application. Though its roots trace back to the mid-20th century (Nadkarni et al., 2011), it almost ceased to progress in the 1960s because it was still more cost-effective to have people process written language instead of machines (Foote, 2019). In the 1980s and 1990s, with more computing power and advances in machine learning, NLP made a resurgence that continues to advance today.

The use of NLP in healthcare is even more recent. An early example is IBM's Watson, first conceptualized in 2006 and initially piloted by two healthcare organizations in 2012 (Best, 2013). Other uses of NLP in healthcare include drug research, imaging diagnostics, predictive diagnostics, and research analysis. Since using NLP has become more relevant in healthcare, attention has been focused on many aspects of unstructured data available in clinical settings. Clinical notes are one common source of unstructured data. While there have been advances and several platforms have emerged to support NLP of clinical notes, it is a topic that still has many challenges to overcome and can be considered to be in its infancy in many ways. Additional research on this topic is essential to continue improving existing methods.

## Objectives and Purpose

While there are many uses for NLP, the focus of this paper was to extract structured meaning from unstructured text with the primary objective of extrapolating high mortality conditions from clinical notes. The extrapolation was done by classifying text related to specific medical conditions, including cancers, diabetes, heart diseases, kidney diseases, and liver diseases. Based on a study published in Jama Network Open, the top five causes of missed diagnoses are colorectal cancer, lung cancer, breast cancer, myocardial infarction (heart attack), and prostate cancer (Schiff et al., 2022). Cancer and heart disease have been the top two causes of death in the US since 1933 (Harding, 2018). The focus on these conditions aims to improve disease identification where the most significant impact can be realized. The process of NLP disease detection could then be expanded upon for more medical conditions or other uses such as medical research, preventative healthcare, or insurance underwriting.

With patients visiting multiple physicians and other medical professionals, pulling together a summary picture that highlights areas of significant concern is critical. Specifically, identifying high mortality conditions can be helpful to healthcare providers in providing improved ongoing and preventative care. Early detection of high mortality conditions can lead to extended patient life and lower costs for providers and patients since medical costs may be higher later in disease progression.

## Research Questions

To address the problems previously identified, over-reliance on structured data, time-consuming labeling of clinical notes, and prominence of highly specialized models, this paper attempted to answer three questions:

* Can high mortality conditions be extracted from clinical notes using NLP?
* Can this be accomplished without excessively labeling the notes using unsupervised learning models?
* Can it be done in a general and scalable way for multiple diseases?

Answering these questions was done by case study analysis comparing and contrasting several models using clinical notes data.

## Significance

The primary benefits of this research are for physicians, mainly primary care providers, and patients. Benefits to primary care providers include reducing time to review data captured by other physicians. Simplifying and improving the accuracy of data aggregation across multiple providers also improves patient care outcomes, a benefit for physicians and patients. The primary benefit to patients is earlier detection of disease. The obvious benefit of this is the extended life of the patient. However, it also has the benefit of lowering costs, as medical intervention earlier in disease progression tends to be less costly.

## Assumptions and Limitations

Any medical data analysis has some limitations to ensure that privacy is protected from moral, ethical, and legal standpoints. For this paper, having de-identified data was a requirement. Using de-identified data meant that while there was a reasonable volume of data to examine, there was not a link at an individual level between encounters. This link would ideally be used to compare information from different encounters to create a fuller picture and find potential comorbidities.

Another limitation was the number of conditions to identify. The focus on high mortality conditions provided a focus and a framework that could, in theory, be used to expand to additional conditions but also imposed limitations on the study. Identifying beneficial conditions to classify is itself a data science problem to solve.

# Chapter 2: Literature Review

While Natural Language Processing has begun to gain traction more recently, its history dates back to the 1940s. This literature review follows a chronological look at NLP history and advancements. There are many uses for NLP, and recent advances have seen the field grow in many directions, such as sentiment analysis, dialogue generation, speech recognition, and classification. Applying NLP to the medical domain is a common undertaking. The organization of this literature review first looks at Natural Language Processing in general, then looks at the application of NLP to healthcare, including more recent applications of NLP to clinical notes.

## NLP Overview and History

It is difficult to pinpoint the true beginning of Natural Language Processing. Much of the literature about the history of NLP cites Alan Turing as the originator. Turing was a mathematician who developed a test in 1950 to determine whether a computer could think (Turing, 1950; Johnson, 2022). Other sources point to a slightly earlier start with a University of Wisconsin - Madison graduate, Warren Weaver, who wrote a memorandum in 1949 later referred to as the Weaver Memorandum (Weaver, 1949; Louis, 2020). This memorandum included suggestions for ways to translate text to elicit meaning. Among the suggestions was a form of neural networks.

There were some advancements in the 1960s and 1970s and notable applications such as ELIZA in 1966. ELIZA was an application created to have a computer mimic one side of a conversation (Louis, 2020). Notably, this was designed to mimic the conversation of a psychiatrist, highlighting the focus of even early NLP processes on medical solutions. Early NLP systems were based mainly on rules devised by hand. Not surprisingly, this long-hand method of developing rules for a machine to comprehend written language was more time-consuming than simply having a person transform the text manually into something more structured. In 1966 a report by the Automatic Language Processing Advisory Committee (ALPAC) noted several reasons advancements in machine translation were unnecessary, including an overabundance of human translators, challenges with the quality of machine translation, and a cost-benefit analysis of using people vs. machines for translation (Hutchins, 1996). The result of this report reduced US government funding for machine translation and is commonly cited as a reason for slowed progression of NLP during this timeframe.

The 1980s and 1990s saw the increase in computing power and the beginning of the move to machine learning and deep learning that helped NLP to expand. The focus on continued advancement endures today. In 2018 Bidirectional Encoder Representations from Transformers (BERT) was introduced by Google (Devlin et al., 2019). BERT is a pre-training NLP technique that predicts language in both directions within a sentence using a masking approach to mask words throughout a sentence. Google began including BERT in its search algorithms in 2019. Since then, additional BERT models have been developed due to the open-source approach used to create BERT, including G-BERT for medical recommendations (Shang et al., 2019).

## NLP Application in Healthcare

NLP and healthcare have a natural connection given the tremendous volume of written information generated and the high level of importance of that information. Translating written healthcare information into meaningful data is critical to patient care, research, and ancillary services such as insurance-related activities. For this reason, it is not surprising that some of the earliest applications of NLP were related to healthcare, such as ELIZA and later IBM Watson.

The advancement of Electronic Health Records (EHR) in the 1990s and subsequent broader internet adoption has also added to the potential for NLP use in EHR. Research using NLP has been used to identify specific diseases, predict the potential for certain conditions, and even extract family history. Clinical notes, in particular, are commonplace for NLP applications. Essentially free-form text with minimal format prescriptions, holding significant detail about a patient and their encounters with physicians.

## Recent NLP Healthcare Studies

Three recent studies demonstrate the value of NLP as applied to healthcare and support the need for further research. These studies also included analyses of both supervised and unsupervised models, which were used as reference points for this case study.

A 2017 study used clinical notes to identify the medical subdomain of the note (Weng et al., 2017). This study compared the results of various types of models, including convolutional recurrent neural networks (CRNN), convolutional neural networks (CNN), support vector machines (SVM), and logistic regression. The results of this study indicated that CNN and CRNN models outperformed SVM and logistical regression models based on comparing the area under the curve (AUC) scores of the models. AUC scores can be used to assess how well a model performs compared to random prediction, with 50% indicating the results for random predictions. The best-performing model had an AUC of over 98%.

Another study completed in 2017 compared rule-based and deep learning approaches for NLP using unstructured data within patients’ discharge summaries (Gehrmann et al., 2017). The purpose of the study was to perform patient phenotyping, which essentially identifies whether a patient has a condition or is at risk of developing a condition. The study found that deep learning approaches, in this case CNN, performed better than various logistic regression and random forest models at classifying patients with or at risk of developing certain diseases.

A 2018 study investigated automatically applying ICD codes to discharge summaries using NLP to determine the diagnoses (Karmakar, 2018). This study included CNNs as well as Recursive Neural Networks (RNN). This study also demonstrated promising outcomes using CNNs. It also supported the Gehrmann study, with very similar results for the CNN models compared to the Gehrmann study. Each study reported F1 scores, a measure used to compare models, of 76% using CNNs.

While all three studies support deep learning and unsupervised learning approaches as viable for medical text interpretation, they also indicate that more research is needed. The history of using NLP for any unstructured data interpretation in healthcare settings is so new that continuing to contribute to the body of knowledge is a critical step in improving and evolving the science.

## Current Limitations and Challenges

NLP has been advancing rapidly recently and has overcome some initial challenges related to cost, efficiency, and quality. An ongoing challenge with NLP for healthcare is the availability of data due to the protected nature of the data being collected. Related to this is another challenge, the notion of specialized vs. generalized NLP. It is common for NLP applications to be very specific in nature and to apply to only a single domain or area of focus. In healthcare settings, this may be attributed to data access limitations. This data accessibility problem continues to make it necessary to essentially reinvent the wheel when applying NLP to a new domain. Finding ways to make NLP more agnostic will improve adoption in new areas.

Bias is also an area where NLP will need to make improvements. Speech patterns, even those that translate to written words, can vary among different cultures. Deep learning algorithms can make mistakes when encountering these differences. Further, the majority of data that can be used for NLP purposes only exists in seven languages, creating a gap in both learning and use of these tools in less-represented populations (Batorsky, 2022).

## Summary

With NLP being a relatively new and expanding field of study, much remains to be learned. Existing models can and will continue to advance. The application of NLP to healthcare remains a prominent and vital area of research. Further, the use of NLP to classify clinical notes, while not novel, has historically focused on specific research. This paper took a different perspective to bring forward the caregiver and patient and the benefits they can realize with NLP applications. Moreover, while the paper focused on the classification of specific high mortality diseases, the premise is that this can be a framework to use for additional disease classification in order to make progress toward more generalized NLP applications.

# Chapter 3: Methodology

The primary objective of this paper was extrapolating high mortality conditions from clinical notes, which had some significant considerations concerning data and methods. The data needed to be robust enough to allow for close-to-real-world text processing while also allowing for the validation of disease prediction. The methods used needed to be able to work specifically with medical data, as a typical NLP library using a standard English vocabulary would likely not have the detail needed to recognize and understand medical terminology.

Access to data for medical analysis is challenging due to the need for privacy with individuals’ medical information. It is common for NLP processes dealing with medical text to be developed for specific research. One reason for this can be attributed to access limitations, where researchers may have restrictions imposed for legal reasons on what data they can use in their research. One of the questions this paper attempted to answer was whether NLP for clinical notes could be done in a more generalized manner rather than for specific research.

Another challenge with NLP for clinical notes is the notion of labeling the notes for the purposes of classification with supervised learning methods. Labeling can be a time-consuming and error-prone task. This paper examined both supervised and unsupervised learning methods in an effort to determine if NLP for clinical notes could be successful without the added labeling step.

## Dataset Overview

The data used in this research was originally from mtsamples.com, which contains approximately 5,000 de-identified transcription reports from 40 different medical specialties. These were subsequently scraped and compiled into a CSV file containing 4,999 transcription records by Tara Boyle and made available on Kaggle (Boyle, 2018). The initial intent was to attempt to classify the dataset based on medical specialty. This paper instead tried to predict disease classification.

The data set contained only five columns. Four columns (description, medical specialty, sample name, and keywords) were taken directly from the mtsamples.com webpage for each transcription. The fifth column, transcription, was compiled from multiple fields that vary across all the transcriptions. Within this field, an indication of the original field label was included in a somewhat consistent format of an all-caps header-like text string.

The primary data fields used in this research are the keywords and transcription fields. The keywords field was taken directly from the mtsamples.com site and contained up to 36 keywords per transcription record, with a median of 13 keywords, though not all transcription records had keywords. Figure 1 shows a distribution of keyword counts per transcription record. In order to define a label that identified diseases for supervised learning models, the keyword field was determined to contain relevant information. The transcription field may have also contained diagnosis information, including disease information; however, it was inconsistently used throughout the transcription records and therefore was determined to be less reliable. It is also important to note that the keywords may have contained information that is not disease related. A discussion about identifying diseases is covered in the disease labeling section.

**Figure 1**  
Count of Keywords per Transcription

Chart, histogram

Description automatically generated

*Note:* Transcriptions with zero keywords are not included.

## Data Cleansing

Since this research focused on natural language processing and the ability to extract useful, structured information from unstructured data, minimal preliminary data cleansing was performed prior to beginning the overall process, which itself included text-related cleansing, to retain the nature of the transcription data. Records that did not contain either any keyword data or any transcription data were removed. One specialty, autopsy, was eliminated entirely due to this. This filtering left 3,898 records for further analysis.

### General Data Cleansing

As previously noted, the transcription field contained a formatted way to segregate each section of the transcription record, which varied across medical specialties. For example, a general medicine transcription may have contained sections including a chief complaint, past medical history, and social history, whereas a surgery transcription may have contained a preoperative diagnosis, a postoperative diagnosis, and techniques used. This segregation was done by adding an all-caps text of the field label followed by the text for that field. The challenge with this format was that the record itself might have all-caps data that was not a header and that was useful for the NLP analysis, so the all-caps text could not be globally removed. Headers were typically followed by a colon and a comma, though this was found to be somewhat inconsistent in some cases containing one or two spaces between the colon and comma. It was important to identify these headers as the header text may have contained medical-like text but was itself not part of the actual transcription and could therefore be mistaken for medical information during the NLP process. The three formats detailed here were used to identify and remove the headers from the transcription text before processing. It should be noted that it is possible that some headers were not successfully removed if they did not follow one of the three formats listed.

### NLP Specific Data Cleansing

When working with unstructured text data and trying to extract meaning from it, there are data cleansing steps, also referred to as text preprocessing, that need to be taken to make the text more understandable to a machine. These preprocessing steps include tokenization, lemmatization, removing stop words and special characters, and vectorization. Many tools used for NLP have their own implementation of these steps that include custom dictionaries and models that can be easily accessed when building NLP models. For this analysis, scikit-learn and scispaCy packages for Python were used. The following is a brief overview of these steps, with additional information about the packages covered in the tools overview section of this paper.

Tokenization is breaking text down into meaningful elements or tokens. This may be done at different levels, such as word tokenization or sentence tokenization. For this analysis, sentence tokenization was used for the transcription field to retain the meaning within sentences, and word tokenization was used for the keyword field. For sentence tokenization, additional information about the words was captured based on their location within the sentence, such as whether it was a verb, adjective, or noun. This becomes important for future preprocessing steps. Tokenization was done using the scispaCy package.

Lemmatization was also done using the scispaCy package. Lemmatization is getting the base form of a word so that its root meaning, or lemma, can be used across a full-text set. When sentence tokenization is used, the additional detail, such as whether the word is a verb or adjective, is also considered when determining the word’s lemma.

Stop words are words that are ultimately insignificant to the underlying meaning of the text. Examples of stop words are "the", "is", and "are". There is no universal list of stop words, but many tools used for NLP include custom dictionaries based on the subject matter for this purpose. In this case, stop words contained in the scispaCy package were used, which includes 326 stop words. Stop words were removed from both the transcription field and the keywords field. In addition to stop words, punctuation and numeric characters were identified as part of the tokenization process and were removed from the transcription field.

Vectorization creates a vector of words within a given collection of text, or corpus. Two common vectorization approaches are count based, also known as bag-of-words, which contains the frequency of words within a given set of text, or doc, within the corpus. The second is TF-IDF (term frequency-inverse document frequency) which uses a combination of the frequency of the term in a doc and weighting based on overall frequency in the corpus. This approach is useful in giving weight to words that may not appear frequently but that can be important to the overall context of the text.

In addition to single-term scoring, n-grams can be incorporated into the vectorization to indicate that strings of ‘n’ number of words should also be considered, not just standalone words. For this paper, 1-, 2- and 3-word n-grams were included. The outcome of vectorization is a matrix of terms, or feature names, and frequencies (for count vectorization) or scores (for TF-IDF vectorization) for each of the docs in the corpus. There are other vectorization approaches that attempt to build upon and improve count and TF-IDF vectorization. However, this paper used TF-IDF vectorization from the scikit-learn package for the transcription field. Figure 2 represents the feature names from the data used in this analysis in a word cloud.

**Figure 2**  
Feature Name Word Cloud

Text

Description automatically generated

Following vectorization, Principal Component Analysis (PCA) was used to reduce multicollinearity within the feature names. This produced a similar matrix with fewer feature names. The initial TF-IDF vector was generated with 1,000 feature names, and the reduced PCA matrix contained 321 feature names.

## NLP Tools Overview

NLP is a complex area of research, and the ability to build upon and reuse applications that have been created and thoroughly researched is important. The selection of tools is an important consideration and can be a substantial aspect of an NLP project. Where in some cases, data cleansing can take up the largest portion of time; for NLP projects finding the right tool to do the data cleansing can be a similar undertaking. This paper used three tools for text analysis: scikit-learn, scispaCy, and the Unified Medical Language System (UMLS). Some others were investigated, and while they may provide value in other NLP projects, these stood out as the best for this particular use case.

### Scikit-learn

Scikit-learn is a Python library used for predictive analytics with supervised and unsupervised learning models. A strength of scikit-learn is its approach to cross-validation and model selection, which is done with fairly minimal coding and simple-to-understand results. For this paper, vectorization, model creation, parameter selection, and model validation were all completed using scikit-learn. TensorFlow is a common alternative to scikit-learn, which is more robust at neural network model generation. While neural network models were created and evaluated for this project, the strengths offered by scikit-learn for other model creation were a guiding factor in its selection.

### ScispaCy and UMLS

ScispaCy is a Python package that contains models specifically designed for working with medical or clinical text and creating NLP models for this domain. It is built upon another package, spaCy, used for NLP text extraction. A common alternative to spaCy is NLTK. At the time of this writing, there was no counterpart for the medical text layer that scispaCy provides, which was ultimately the deciding factor in text extraction tools. In addition to custom models, scispaCy provides a linker that connects directly to the Unified Medical Language System (UMLS). The UMLS is a standardized set of medical vocabularies that provides access to millions of medical terms, concepts, and relationships that combine information from many other medical vocabularies (Bodenreider, 2004). The UMLS linker in scispaCy was used to develop a mapping of keywords to diseases. In addition, scispaCy models were used for lemmatization and stop word removal. This allowed for medical domain-specific terminology to be incorporated into these steps.

## Disease Labelling

In order to conduct supervised classification modeling to predict a disease, a suitable disease label needed to be assigned to the transcriptions. The keyword field was designated as the initial source of this label; however, it needed to be standardized. To accomplish this, the keywords were lemmatized, and then a scispaCy model linked to the UMLS was used to retrieve linked entities using an 80% threshold for entity matching. These entities were filtered to those that contained a Semantic Type Unique Identifier (TUI) that corresponded to a disease or syndrome (T047) or a neoplastic process (T191) (Bodenreider, 2004). The neoplastic process results contained cancer-related and non-cancer-related results that were further filtered.

The initial set of keywords contained 10,327 unique keywords. Linking these keywords to standard diseases resulted in 2,017 unique disease labels. In order to focus on the five specific high mortality conditions selected for this analysis, this data needed to be further aggregated to combine heart-related diseases, kidney-related diseases, liver-related diseases, cancers, and diabetic diseases. 241 of the 2,017 disease labels were mapped to one of these five categories. This step was performed manually. The manual mapping results can be found in Appendix 1. This step highlights one of the challenges with NLP, which is labeling. In order to perform supervised learning, a label is needed. However, with unstructured data, this can be time-consuming and difficult to obtain reliably. For this reason, both supervised and unsupervised learning approaches were used in order to compare results across these different types of models.

The final step in the labeling process was to apply the aggregated label to the transcription records. The data was also further filtered to only include transcription records that were assigned at least one of the five aggregated disease labels. This resulted in 1,173 records to use for further analysis. Table 1 shows the final distribution of the data used for modeling by medical specialty as well as the original distribution for reference. Six medical specialties were excluded from the final modeling: Allergy / Immunology, Autopsy, IME-QME-Work Comp etc., Psychiatry / Psychology, Sleep Medicine, and Speech – Language.

**Table 1**  
Medical Specialty Distribution

Table

Description automatically generated

The distribution of transcription records across the five aggregated disease labels is shown in Figure 3. There was an imbalance between cancer records and all other diseases. This was taken into account when modeling to use balanced approaches. Additionally, a binary approach was explored using only cancer compared to non-cancer disease classification to improve balance. The distribution of transcription records between cancer and non-cancer is shown in Figure 4.

**Figure 3** *High Mortality Disease Label Distribution*

***Chart, bar chart

Description automatically generated***

**Figure 4** *Cancer/Non-Cancer Label Distribution*

Chart, bar chart

Description automatically generated

## Classification Modeling Approach

Research shows that naïve Bayes and decision trees are very common models for NLP classification. This paper analyzed these types of models in addition to logistic regression and neural networks. These selections primarily aimed to compare results across the different approaches, including comparisons between supervised and unsupervised learning models. While supervised learning models, such as naïve Bayes and decision trees, have commonly been used for NLP modeling, unsupervised learning has been gaining traction. It provides an enormous benefit by eliminating the need to label training data which can be a time-intensive task.

As previously noted, the multiclass dataset had a significant imbalance, with cancers comprising 66% of all labels. While there are ways to account for imbalance, since the balance between cancer and non-cancer labels was less imbalanced, with all other labels combined as non-cancer comprising 34% of the dataset, binary classification models were also created.

### Logistic Regression

Logistic regression is a discriminative classification approach and is a typical starting point for classification modeling. For this analysis, it was a useful comparison point to other models. Logistic regression is inherently a binary classification approach; however, a One-vs-Rest (OvR) approach can be used, which treats each classifier as a binary classifier, comparing it to the "rest" of the classifiers. For the multiclass classification, OvR was used in the Logistic Regression model.

One of the drawbacks of logistic regression is its limited ability to understand multiple relationships. In a real-world setting, it is very likely that clinical notes would not resolve to a single binary decision. In other words, one clinical note may contain indications of multiple diseases. There are ways to account for this, such as with a multinomial approach instead of the OvR approach. However, it begins to make the model more complex when one of the benefits of the logistic model is its simplicity, and other models may inherently be better suited to this scenario.

### Naïve Bayes

Naïve Bayes is a common supervised learning model used for NLP that is a simplification of the Bayes theorem that assumes feature independence. The Bayes theorem is used to determine the probability of an event (A) given known conditions (B), as shown in the following equation.

Naïve Bayes adds an assumption that A and B are independent. For NLP purposes, A and B are the words, including the 1-, 2- and 3-word n-grams, selected to be the features. Since the order of the features is not particularly important, we can assume independence, at least for the purposes of testing the model. If it performs poorly, that may be an indication of dependence we are overlooking.

There are several naïve Bayes methods, including Gaussian, the typical or classic use of naïve Bayes, as well as multinomial, complement, Bernoulli, and categorical. This analysis used the Gaussian method, which was suited for the continuous multiclass features (when using the TF-IDF scores for each feature) used in this analysis. Multinomial and complement methods are suitable for multi-label classification, Bernoulli is suitable for binary classification, and categorical is suitable for categorical features.

### Decision Trees

Decision trees are another form of supervised learning that relies on piecewise decisions that, when visualized, resemble a tree with nodes established at decision points creating branches from each node, similar to a flowchart. The final node, where no further decisions can occur, is a leaf node. Pruning is used to reduce decision nodes to the most impactful ones to prevent overfitting. There are many ways to set up a decision tree by determining the maximum depth, leaves, splits, and more. One thing to be mindful of with decision trees is that they can become complex quickly. While outcomes may appear to show better performance with more complexity, caution must be taken to avoid overfitting.

Another form of decision trees is random forest, which combines many decision trees to develop an aggregate of all of them into one model. They are useful when there are many features going into the decision-making, and since text data can produce a high volume of features, they are commonly used for NLP modeling. In this case, only the basic decision tree model was used instead of random forests, as the documents were not overly large, and the feature set was a reasonable size. In some small tests, results tended to be similar between the two approaches, and it is beneficial to use the simpler model whenever possible.

### Neural Networks

The final type of modeling analyzed was neural networks. Neural networks are a form of unsupervised learning, which makes them particularly appealing for NLP tasks where labeling data can be time-consuming and cumbersome. Neural networks are often described as a way to mimic what the human brain does when it makes decisions. Hidden layers in a neural network are made up of nodes where decisions are made and sent to another node before arriving at a final output. One of the challenges of neural networks is the hidden layers which can make it difficult, if not impossible, to explain how the solution was generated, referred to as a black box solution. Figure 5 shows a diagram example of a neural network.

**Figure 5**  
Neural Network Example

A picture containing text, businesscard, vector graphics

Description automatically generated

*Note:* Graphic by [TseKiChun](https://commons.wikimedia.org/w/index.php?title=User:TseKiChun&action=edit&redlink=1), distributed under a [CC BY-SA 4.0 license](http://creativecommons.org/licenses/by/2.0/).

There are several different types of neural networks, including Multi-layer Perceptron (MLP), which is what was used for this analysis, Convolutional Neural Networks (CNN), and Recursive Neural Networks (RNN). While CNN has shown promise for NLP (Assale et al., 2019), scikit-learn only includes MLP methods, which was the reason for the selection. The primary difference is that in MLP models, all nodes connect to each other, whereas in CNN, the nodes may also be pooled and not directly connected to one another. MLP models are easier to train, whereas CNN models are preferred for image classification though they have also performed well on non-image data (Brownlee, 2022).

## Model Testing and Verification Approach

One of the benefits of using scikit-learn for NLP modeling is the relative ease of tuning hyperparameters and retrieving verification metrics. A variety of metrics are available that lend themselves to handling multiclass scenarios in order to derive averages, including the type of average to use, such as weighted, micro, or macro. As this analysis looked at both multiclass and binary classification, different approaches were used for each but generally aligned across models within each approach to make comparisons.

For the purposes of training and testing the models, the data was split to have 75% of the data used for training and 25% used for testing. The data was also stratified during this split using the disease labels. This ensured the training and test data each contained distributions of each disease label that were consistent with the full dataset. This avoids the potential for all data with a certain label to be grouped into only the training or only the test data set, with no data available for the alternate activity.

### Parameter Tuning

Two approaches were used to tune parameters that allowed for hyperparameter tuning using scikit-learn, Grid Search CV, which evaluates all combinations of hyperparameters defined, and Randomized Search CV, which evaluates a random set of hyperparameters defined. In some initial testing, results were fairly similar regardless of the parameters used, and some selections required significant time to execute. A randomized search was completed in these cases, which included logistic regression, decision trees, and neural networks. A grid search was conducted for naïve Bayes as the parameter options were much smaller, and this model tended to have worse results than the other models. Ensuring all options were being evaluated meant a potentially better solution was not missed with a randomized search. An overview of the parameters selected for each model is shown in Table 2. All hyperparameter tuning was completed with 5-fold cross-validation. Randomized searches sampled ten random parameter settings.

**Table 2**  
Hyperparameter Tuning Summary

Graphical user interface, text

Description automatically generated

*Note:* Not all hyperparameter combinations are valid.

While all the models returned good results with the best parameters selected during tuning, there were some considerations for the final parameters used. The logistic regression model needed a significant number of iterations to converge, which could indicate instability in the model. This model was primarily included for reference purposes. The decision tree model performed quite well but tended to have a high depth to reach the optimal result. The max depth was restricted in the parameter tuning to a random integer between 2 and 10. Additionally, larger datasets tend to work better for decision tree purposes. Smaller datasets may lead to somewhat unstable results, as small changes in small datasets can have a more significant effect on the overall tree.

### Classification Metrics

Common metrics for classification problems are accuracy, precision, recall, and F1-score. Accuracy is the proportion of correct predictions out of all of the predictions. Precision is the proportion of true-positive predictions in all positive predictions (true-positives and false-positives). In contrast, recall is the proportion of true-positive predictions in all actual positives (true-positives and false-negatives). And the F1-score is the harmonic mean of precision and recall and is used to compare these two metrics across different classifiers. While accuracy is most commonly referenced, precision and recall can be more important in certain circumstances. In medical classification, recall is considered the more important metric because it minimizes false negatives. A false negative in medical situations can lead to missing a diagnosis which can have particularly detrimental consequences. For this reason, recall was used as the scoring metric to maximize during hyperparameter tuning. For multiclass models, a weighted average recall was used, which combined the recall for all of the classes with weighting to account for the imbalance present in the dataset.

## Summary

The methods used to analyze the data revealed additional information about the state of the dataset, including its strengths and weaknesses. While there were some limitations with the data being used for this analysis, namely that the labeled dataset was relatively small and there were imbalances with the classifiers, the modeling results still demonstrated that benefits could be achieved through NLP for clinical notes. The limited nature of the dataset demonstrates one of the challenges with NLP for medical data in that obtaining de-identified data is difficult, and what is obtainable generally will have some type of limitation. The next chapter will demonstrate the value that can be achieved with NLP for clinical notes, even though larger datasets are needed to continue this type of research.

# Chapter 4: Results and Findings

This section will review some high-level descriptive statistics followed by an overview of the results of supervised and unsupervised learning approaches and how the results answer the research questions. Lastly, the results of all models will be compared. The results supported answers to the three questions this paper attempts to answer: can high mortality conditions be extracted from clinical notes, can it be done without labeling the data, and can the solution scale?

## Descriptive Statistics

Word count is a common basic measure in NLP analysis. A 2021 study analyzing average lengths of clinical notes found that the median word count of clinical notes in 2009 was 401 words, and in 2018 it was 642 words (Rule et al., 2021). While dates were not included in the clinical notes used in this analysis, these can be used as a reference point. The median word count of clinical notes in this analysis was 383 words, which is lower than the 2009 median observed in the study. Notes labeled as cancers had the highest median of 403 words, while diabetes had the lowest median of 350 words. The median word count for all disease labels can be found in Figure 6. Figure 7 shows the word count by transcription count.

**Figure 6**  
Median Transcription Word Count by Disease

Chart, bar chart

Description automatically generated

**Figure 7**  
Word Count in Transcriptions

Chart, histogram

Description automatically generated

As described in the previous chapter, TF-IDF was used to select feature names for use in the modeling. Figure 8 shows the top 20 feature names selected for each of the five target diseases. This was determined using the TF-IDF scores, which were summed across all documents for all 1,000 initial feature names selected by the TF-IDF model. While the final number of feature names was reduced using PCA, this gives an overview of the differences already apparent with the transcription text for each of the five target diseases.

**Figure 8**  
Top 20 Feature Names by Disease

Chart, table, bar chart

Description automatically generated

*Note:* Displays the sum of TF-IDF Scores by feature names. Scales vary by disease.

## Supervised Learning Results

In answering the question of whether high mortality conditions can be extracted from clinical notes, accuracy, balanced accuracy, and weighted average recall scores for each of the supervised learning models were reviewed. While model accuracy for multiclass classification models was high for all supervised models, ranging from 80% - 88% without considering the data imbalance, the balanced accuracy was quite a bit lower, between 62% and 76%. Accuracy over 70% can be considered good, and the logistic regression and decision tree models each achieved this at 76% and 75% balanced accuracy, respectively.

Further, when considering recall scores, the outcomes were even more promising. As previously noted, recall can be considered more important for medical-related prediction models to reduce the potential for false negatives. All supervised models achieved a weighted average recall of 80% - 88%. This metric already considers the data imbalance with the weighting. This was a very good result for the multiclass models. Logistic regression and decision trees had the highest weighted average recall at 88% and 84%, respectively. The naïve Bayes model average weighted recall was 80%, which is still quite high and a very good result.

Given that the data was imbalanced, with 66% of clinical notes labeled as cancers, a binary approach was also reviewed to achieve more balance with each class. Binary methods for the supervised models produced much better results, with balanced accuracy scores ranging from 81% - 91%. Logistic regression and decision trees again performed the best. However, the naïve Bayes model achieved an 81% balanced accuracy. This was an increase of 19% over the multiclass naïve Bayes model.

These results indicate that high mortality diseases can be extrapolated from clinical notes using supervised learning with a high level of confidence. Figure 9 shows a comparison of the metrics reviewed for the supervised learning models.

**Figure 9**  
Supervised Learning Model Metrics

Chart, bar chart

Description automatically generated

In answering whether the results are scalable, the individual disease classification results of the multiclass models were reviewed. These demonstrated mixed results, likely attributable to the varied volume of data across the different disease classes. The individual classification outcomes were better when there was more data for a disease label. Conversely, less data led to a worse result. Using the logistic regression multiclass outcome as an example, recall was 95% for cancers and 88% for heart diseases. These are good results, but transcriptions with these disease labels have the highest volume of data at 66% and 16% of the data, respectively. The recall for kidney diseases was only 52%, with kidney diseases making up only 7% of the data. The precision, recall, and F-1 scores for the logistic regression multiclass model are shown in Table 3. The confusion matrices are shown in Figure 9. These demonstrate the count disparities as well as the disparities as a reflection of the percentage of true predictions. Kidney diseases were predicted to be liver diseases nearly as often as they were accurately predicted to be kidney diseases. These are important to take into consideration as the overall model metrics can dilute this detail.

**Table 3**  
Multiclass Logistic Regression Classification Report

Table

Description automatically generated

**Figure 10**  
Multiclass Logistic Regression Confusion Matrices

Calendar

Description automatically generated with medium confidence Chart

Description automatically generated

The binary approach can reduce this deficiency and may replicate a multi-label approach, where a binary outcome per disease is determined. In this case, the binary decision was cancers or non-cancers. Using logistic regression again as an example, Table 4 shows the recall for cancers was 94% and for non-cancers was 87%. This was a much better outcome than the multiclass classification. Figure 11 shows the confusion matrices for the binary logistic regression model. The results for all models tested, including classification reports and confusion matrices, can be found in Appendix B.

**Table 4**  
Binary Logistic Regression Classification Report

Table

Description automatically generated

**Figure 11**  
Binary Logistic Regression Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

## Unsupervised Learning Results

As noted previously, labeling transcription records can be a time-intensive process and can be somewhat unreliable. Different people may label things using different approaches and identify different aspects of the transcription as important; ultimately, labels may be missed or inaccurate.

One of the main questions this paper attempted to answer was whether unsupervised learning models could produce reliable results in order to alleviate the need to label data to create a model. A neural networks model was developed to answer this question, and accuracy and recall were reviewed. The training metrics were also reviewed as part of the model evaluation since labels were not used to train the model. These metrics were evaluated independently and compared to the supervised learning model results.

The training metrics demonstrated a high degree of accuracy and recall, which is to be expected. For the multiclass model, the accuracy and weighted recall were both 94%, and the balanced accuracy was 86%. Given that these were the training results, this shows there was some dilution of the accuracy due to the variance in accuracy scores of each independent class. Looking at the individual classifications, liver diseases had a recall of only 62%, which is quite low for a training model. Table 5 shows the classification report of the neural networks training model.

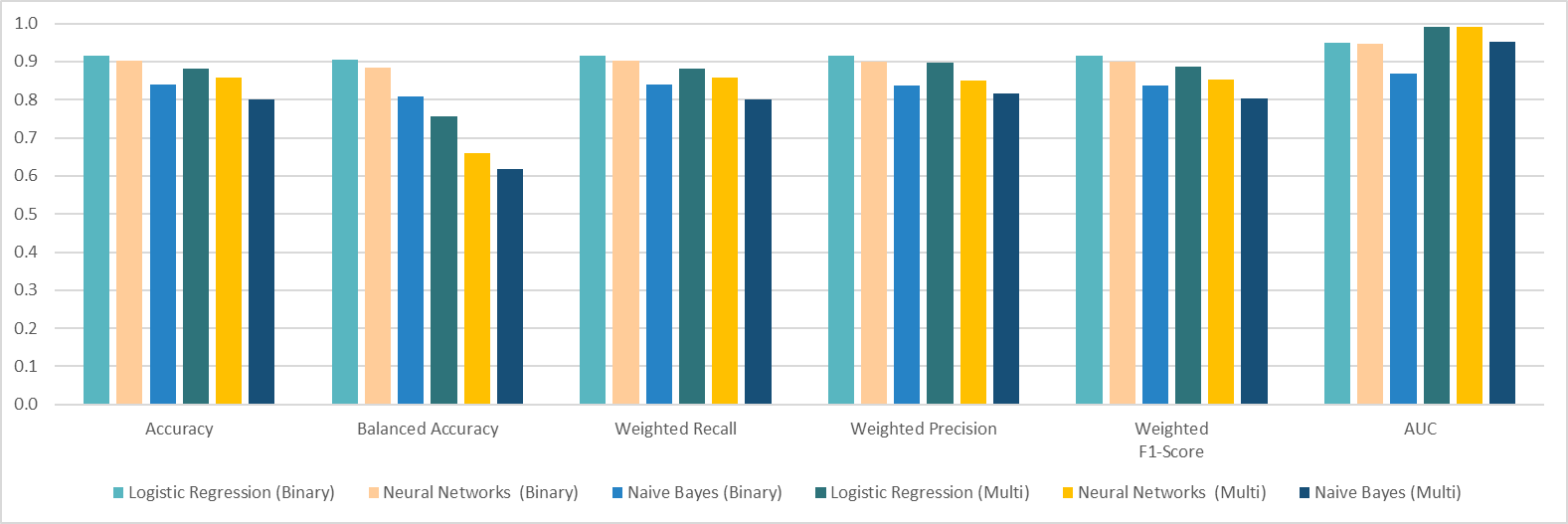
**Table 5**  
Multiclass Neural Networks Training Classification Report

Table

Description automatically generated

The test outcomes for the multiclass model were 66% balanced accuracy and 86% weighted recall. Though these results were not as strong as the logistic regression and decision trees multiclass models, they were better than the naïve Bayes multiclass model by over 4% for balanced accuracy and nearly 6% for weighted recall. The recall scores at the class level were somewhat similar to the logistic regression model. The recall for cancer and heart diseases, the classes with the most data, were the same for each model, 95% for cancer and 88% for heart disease. The recall scores were lower in the neural networks model for diabetes, kidney diseases, and liver diseases. This supports the supervised model findings that more data leads to more reliable results and adds a level of reliability to the neural network model. The model performed equitably to the supervised models where there was enough data. Figure 12 shows the metrics reviewed for the neural network models compared to the logistic regression and naïve Bayes models. These were the best and worst supervised models for comparison.

**Figure 12**  
Unsupervised Compared to Supervised Learning Model Metrics



A binary approach was also reviewed for the neural networks model, and results were generally similar compared to the supervised models. The training model performed better, with 98% balanced accuracy, weighted recall, and 98% recall for both classes, cancers and non-non-cancers. The test data results were again comparable to the logistic regression binary model, with 94% recall for cancers, the same as the logistic regression model, and 83% recall for the non-cancers, only 4% lower than the logistic regression model result. For the overall model performance, it achieved 88% balanced accuracy, more than 7% better than the naïve Bayes binary model and only 2% worse than the logistic regression binary model. The weighted recall was 90%, which is an excellent result in its own right. Comparing it to the supervised models, this was again better than the naïve Bayes model and only 1.4% worse than the logistic regression model. Looking at it another way, out of 294 predictions, the neural networks model incorrectly predicted only four more transcriptions than the logistic regression model. These results strongly support that unsupervised models can be used with high confidence for NLP of clinical notes.

## Comparison of Results

The accuracy, precision, recall, F-1 score, and AUC score metrics for all models tested are found in Table 6. Binary logistic regression performed best across all models, while multiclass naïve Bayes performed the worst. The variance between the best and worst models in each group is approximately 8% for all measures except balanced accuracy and AUC score. The balanced accuracy for the naïve Bayes and neural networks multiclass models is below 70%, indicating less reliability of the models.

**Table 6**  
All Model Results

Table

Description automatically generated

The ROC Curve and AUC scores for the binary models are shown in Figure 13. This indicates that all models produced results better than random predictions, with similar curves for logistic regression, decision trees, and neural networks. Naïve Bays had a lower AUC, but the score of 0.868 still indicated high confidence in predictions from this model.

**Figure 13**  
ROC Curve and AUC Scores for Binary Models

Chart

Description automatically generated

## Summary

The primary observation made as a result of this analysis was that while unsupervised learning models may not perform as well as supervised learning models, they still performed very well. The tradeoff in time saved with eliminating the need to label data to train a model may be acceptable given the high confidence unsupervised models provide. Additionally, while binary models performed better than multiclass models, this may be attributed more to the imbalance of the data than the models. Furthermore, in a real-world scenario, any clinical note may have any number of diseases detected. For this reason, a multi-label approach needs to be investigated and can be thought of as multiple binary decisions being made on each transcription.

# Chapter 5: Discussion

With the implications of costs of missed and late diagnoses and the increasing volume of healthcare data, particularly unstructured healthcare data, this case study aimed to answer three questions:

* Can high mortality conditions be extracted from clinical notes using NLP?
* Can this be accomplished without excessively labeling the notes using unsupervised learning models?
* Can it be done in a general and scalable way for multiple diseases?

This section will cover the answers that were found to each of these questions based on the analysis completed and the importance of the answers. Limitations of the case study and future considerations will also be discussed.

## Extracting High Mortality Conditions from Clinical Notes

This case study demonstrates that extracting high mortality conditions from clinical notes is feasible with a high degree of accuracy. Isolating prediction of cancer as a binary classification problem, all models tested resulted in recall and precision of between 84% and 92%. This demonstrates that disease classification can be achieved in a way that minimizes false negatives as well as false positives. Furthermore, while detecting a single disease produced better outcomes, detecting multiple diseases with a multiclass approach also produced results that warrant further consideration. Multiclass approaches resulted in precision and recall of between 80% and 90%, though the imbalance of data is likely skewing this due to the high precision and recall of cancer. The results of individual disease classification indicate that more data would produce better results for multiclass classification.

## Unsupervised Learning Models for NLP of Clinical Notes

Labeling data in order to train an NLP model is a time-consuming activity. While this action does result in better model outcomes using supervised learning, the decline in model accuracy when using unsupervised learning is fairly low. There was only a 2% difference in balanced accuracy between the best supervised learning model and the only unsupervised learning model for binary approaches. The difference in recall and precision between these models was also only 2%. It is worth considering the tradeoff in time saved to achieve nearly identical results. Additionally, more research is needed into other unsupervised learning approaches that may result in even better outcomes. Based on the 2017 Gehrmann study referenced in the literature review of CNNs compared to logistic regression and random forest models, CNNS, an unsupervised learning method, performed better than the supervised learning models (Gehrmann et al., 2017). The Gehrmann study included cancer and heart disease, which were conditions also analyzed in this case study. This supports the need for further research into unsupervised learning models for prediction of high mortality diseases.

## Scalability of NLP for Clinical Notes

Scalability is the most challenging outcome to measure. While the study was not based on a specific disease and, in that regard, could be considered agnostic to disease classification, there was not enough data for any disease other than cancer to fully support scalability. Clinical notes labeled as heart disease had the second highest volume of records and performed generally well in the multiclass modeling, which supports some level of scalability. On the best performing multiclass model, logistic regression, the F1-score for cancer was 96%, while for heart disease, it was 88%. For the worst performing model, naïve Bayes, the F1-score for cancer was 90%, and for heart disease, it was 80%. This suggests that with more balance in the data, a more balanced outcome may be achieved that retains a high level of precision and recall for multiple diseases.

## Case Study Limitations

The biggest limitation of this case study was the volume of data. More data would have been desirable, but as previously noted, this demonstrates one of the overall challenges with medical data research which is access to de-identified data. That said, I believe the results are still reliable enough to support the answers to the questions posed. The data was not pre-selected for certain diseases, so the outcome was not known prior to the initiation of the case study.

Connected to low data volume is another limitation of the imbalance of disease labels. Cancer was the most significant of the diseases analyzed, and not surprisingly produced the best results. Models were adjusted to account for the imbalance, including binary classification modeling. However, the imbalance does leave more room for research into additional disease identification.

## Recommendations for Future Research

As previously noted, future research should be focused on additional disease classification as well as more in-depth research toward unsupervised learning methods. While scikit-learn was limited to MLP neural networks, other tools can be explored that support a wider range of unsupervised learning, such as TensorFlow.

Another area that requires additional research is multi-label classification. This would ensure that disease identification would not be limited to a single disease. In line with this, cross-referencing of multiple clinical notes for a single patient would provide a greater foundation for disease identification. As each identification could be considered a binary decision, the research supporting binary classification shows promise for this area of research.

# Conclusion

With billions of dollars being spent and tens of thousands of lives being lost to misdiagnoses each year, the importance of early disease identification can’t be overstated. The increasing reliance on segregated medical support and the volume of unstructured data brings resources like clinical notes to the forefront of potential areas to find solutions. Finding ways to expand the use of NLP to more areas of medicine is a critical need.

This case study reviewed eight different approaches to NLP for clinical notes to detect five high mortality conditions. All eight models produced results that extend insights into the type of data needed for further research. Several of the models produced results with accuracy and recall metrics over 80%. Even with the limitations identified, this case study has demonstrated that using NLP for earlier and more reliable prediction of high mortality diseases can provide significant benefits to physicians and patients. Reducing the time needed by physicians leading to earlier disease identification means better patient outcomes, lower loss of life, and reduced costs for all involved.

With previous studies using NLP for disease detection also demonstrating the success that can be achieved, this case study further supports those findings. It adds to previous research by demonstrating disease detection from unstructured text can be done more generally. And it further supports previous research demonstrating that unsupervised approaches can be very successful, requiring less time to develop solutions. Continuously improving the reliability and expanding the application of NLP models for medical use can and will save time, save money, and, most importantly, save lives.

# References

Al-Aidaroos, K. M., Bakar, A. A., & Othman, Z. (2012). Medical Data Classification with Naive Bayes Approach. *Information Technology Journal*, *11*(9), 1166–1174. <https://doi.org/10.3923/itj.2012.1166.1174>

Assale, M., Dui, L. G., Cina, A., Seveso, A., & Cabitza, F. (2019). The Revival of the Notes Field: Leveraging the Unstructured Content in Electronic Health Records. *Frontiers in Medicine*, *6*. <https://www.frontiersin.org/articles/10.3389/fmed.2019.00066>

Batorsky, B. (2022, April 9). *New Technology, Old Problems: The Missing Voices in Natural Language Processing*. The Gradient. <https://thegradient.pub/nlp-new-old/>

Best, J. (2013, September 9). IBM Watson: The inside story of how the Jeopardy-winning supercomputer was born and what it wants to do next. TechRepublic. <https://www.techrepublic.com/article/ibm-watson-the-inside-story-of-how-the-jeopardy-winning-supercomputer-was-born-and-what-it-wants-to-do-next/>

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(90001), 267D – 270. <https://doi.org/10.1093/nar/gkh061>

Boyle, T. (2018, October 15). *Medical Transcriptions*. <https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions>

Brownlee, J. (2022, August 11). When to Use MLP, CNN, and RNN Neural Networks. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/when-to-use-mlp-cnn-and-rnn-neural-networks/>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <http://arxiv.org/abs/1810.04805>

Divakaran, M. (2019, September 2). ScispaCy for Bio-medical Named Entity Recognition(NER). *Medium*. <https://maheshdivakaran.medium.com/scispacy-for-bio-medical-named-entity-recognition-ner-63ed548f1df0>

Foote, K. D. (2019, June 17). *A brief history of natural language processing (NLP)*. DATAVERSITY. <https://www.dataversity.net/a-brief-history-of-natural-language-processing-nlp/>

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote Jr., J., Moseley, E. T., Grant, D. W., Tyler, P. D., & Celi, L. A. (2017, March 25). *Comparing rule-based and deep learning models for patient phenotyping*. [1703.08705] <https://arxiv-export-lb.library.cornell.edu/abs/1703.08705?context=cs.NE>

Harding, M. C. (2018). Transitions From Heart Disease to Cancer as the Leading Cause of Death in US States, 1999–2016. *Preventing Chronic Disease*, *15*. <https://doi.org/10.5888/pcd15.180151>

Health, D. (2018, August 17). Understanding clinical notes. Medium. <https://medium.com/@dothealth/understanding-clinical-notes-d98afeabf90>

HealthITSecurity. (2022, January 14). *What Are the Benefits of Natural Language Processing Technology?* HealthITAnalytics. <https://healthitanalytics.com/features/what-are-the-benefits-of-natural-language-processing-technology>

*How Misdiagnosis Is Increasing Healthcare Costs and How MORE Health is Fixing the Problem*. (n.d.). Retrieved November 30, 2022, from <https://www.morehealth.com/blog/how-misdiagnosis-is-increasing-healthcare-costs-and-how-more-health-is-fixing-the-problem>

Hutchins, J. (1996). ALPAC: The (In)Famous Report. MT News International, pp. 9–12. <https://pangeanic.co.uk/wp-content/uploads/sites/2/2014/04/ALPAC-1996.pdf>

Johnson, S. (2022, March 7). The Turing test: AI still hasn’t passed the “imitation game.” *Big Think*. <https://bigthink.com/the-future/turing-test-imitation-game/>

Karmakar, A. (2018). *Classifying medical notes into standard disease codes using Machine Learning* (arXiv:1802.00382; Version 1). arXiv. <http://arxiv.org/abs/1802.00382>

Kavuluru, R., Han, S., & Harris, D. (2013). Unsupervised Extraction of Diagnosis Codes from EMRs Using Knowledge-Based and Extractive Text Summarization Techniques. *Advances in Artificial Intelligence. Canadian Society for Computational Studies of Intelligence. Conference*, *7884*, 77–88. <https://doi.org/10.1007/978-3-642-38457-8_7>

Louis, A. (2020, July 7). A Brief History of Natural Language Processing—Part 1. *Medium*. <https://medium.com/@antoine.louis/a-brief-history-of-natural-language-processing-part-1-ffbcb937ebce>

Medicine, I. of. (2012). *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America*. <https://doi.org/10.17226/13444>

Monsters, D. (2017, September 26). 7 types of Artificial Neural Networks for Natural Language Processing. *Medium*. <https://medium.com/@datamonsters/artificial-neural-networks-for-natural-language-processing-part-1-64ca9ebfa3b2>

Mulyar, A., Schumacher, E., Rouhizadeh, M., & Dredze, M. (2020). Phenotyping of Clinical Notes with Improved Document Classification Models Using Contextualized Neural Language Models (arXiv:1910.13664). arXiv. <http://arxiv.org/abs/1910.13664>

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: An introduction. Journal of the American Medical Informatics Association, 18(5), 544–551. <https://doi.org/10.1136/amiajnl-2011-000464>

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

PinnacleCare. (2016). *The Human Cost and Financial Impact of Misdiagnosis* [White paper]. [Human-Cost-Financial-Impact-Whitepaper.pdf (pinnaclecare.com)](https://www.pinnaclecare.com/forms/download/Human-Cost-Financial-Impact-Whitepaper.pdf)

RBC Capital Markets | Navigating the Changing Face of Healthcare Episode. (n.d.). Retrieved October 15, 2022, from <https://www.rbccm.com/en/gib/healthcare/story.page>

Rule, A., Bedrick, S., Chiang, M. F., & Hribar, M. R. (2021). Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center. *JAMA Network Open*, *4*(7). <https://doi.org/10.1001/jamanetworkopen.2021.15334>

Savage Law Firm. (2014, May 8). Top 5 Causes of Misdiagnoses and Diagnostic Errors. *Savage Law Firm*. <https://savagelawyer.com/top-5-causes-of-misdiagnoses-and-diagnostic-errors/>

Schiff, G. D., Volodarskaya, M., Ruan, E., Lim, A., Wright, A., Singh, H., & Reyes Nieva, H. (2022). Characteristics of Disease-Specific and Generic Diagnostic Pitfalls: A Qualitative Study. *JAMA Network Open*, *5*(1), e2144531. <https://doi.org/10.1001/jamanetworkopen.2021.44531>

Shang, J., Ma, T., Xiao, C., & Sun, J. (2019). Pre-training of Graph Augmented Transformers for Medication Recommendation. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 5953–5959. <https://doi.org/10.24963/ijcai.2019/825>

Shobha, G., & Rangaswamy, S. (2018). Chapter 8—Machine Learning. In V. N. Gudivada & C. R. Rao (Eds.), *Handbook of Statistics* (Vol. 38, pp. 197–228). Elsevier. <https://doi.org/10.1016/bs.host.2018.07.004>

Solomon, M. D., Tabada, G., Allen, A., Sung, S. H., & Go, A. S. (2021). Large-scale identification of aortic stenosis and its severity using natural language processing on electronic health records. *Cardiovascular Digital Health Journal*, *2*(3), 156–163. <https://doi.org/10.1016/j.cvdhj.2021.03.003>

Solutions, G. (2022, February 16). *How to Deal with Unstructured Healthcare Data*. <https://gaine.com/blog/mdm/how-to-deal-with-unstructured-healthcare-data/>

TseKiChun. (2021). *English: Neural network explain*. Own work. <https://commons.wikimedia.org/wiki/File:Neural_network_explain.png>

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. <https://doi.org/10.1093/mind/LIX.236.433>

Velupillai, S., Mowery, D., South, B. R., Kvist, M., & Dalianis, H. (2015). Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. *Yearbook of Medical Informatics*, *24*(01), 183–193. <https://doi.org/10.15265/IY-2015-009>

Weaver, W. (1952). Translation. *Proceedings of the Conference on Mechanical Translation*. EarlyMT 1952, Massachusetts Institute of Technology. <https://aclanthology.org/1952.earlymt-1.1>

Weng, W.-H., Wagholikar, K. B., McCray, A. T., Szolovits, P., & Chueh, H. C. (2017). Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC Medical Informatics and Decision Making, 17(1), 155. <https://doi.org/10.1186/s12911-017-0556-8>

# Appendix A

The following is the mapping of disease data obtained from scispaCy and UMLS to an aggregated label indicating cancers, diabetes, heart diseases, kidney diseases, and liver diseases. The percentage represents the overall dataset percentage.

**Table A1**   
Aggregated Disease Mapping

| **scispaCy/UMLS Value** | **Count** | **%** | **Aggregated** |
| --- | --- | --- | --- |
| cervical cancer | 142 | 4.6% | Cancers |
| Kidney Failure | 115 | 3.7% | Kidney Diseases |
| Diabetes Mellitus | 81 | 2.6% | Diabetes |
| Anal carcinoma | 74 | 2.4% | Cancers |
| Carcinoma of anal margin | 73 | 2.3% | Cancers |
| Carcinoma | 68 | 2.2% | Cancers |
| Coronary Arteriosclerosis | 62 | 2.0% | Heart Diseases |
| Gallbladder Well Differentiated Endocrine Tumor/Carcinoma | 59 | 1.9% | Cancers |
| Lip and Oral Cavity Carcinoma | 55 | 1.8% | Cancers |
| Oral Cavity Carcinoma | 55 | 1.8% | Cancers |
| Heart failure | 51 | 1.6% | Heart Diseases |
| Coronary Artery Disease | 48 | 1.5% | Heart Diseases |
| Coronary heart disease | 46 | 1.5% | Heart Diseases |
| Renal Cell Carcinoma | 45 | 1.4% | Cancers |
| Childhood Acute Lymphoblastic Leukemia | 44 | 1.4% | Cancers |
| Childhood Soft Tissue Sarcoma | 44 | 1.4% | Cancers |
| Atrial Fibrillation | 40 | 1.3% | Heart Diseases |
| Chronic kidney disease stage 5 | 37 | 1.2% | Kidney Diseases |
| Oropharyngeal Carcinoma | 37 | 1.2% | Cancers |
| Adenocarcinoma | 36 | 1.2% | Cancers |
| Cervical Adenocarcinoma | 33 | 1.1% | Cancers |
| Ductal Carcinoma | 33 | 1.1% | Cancers |
| Adenocarcinoma of anus | 32 | 1.0% | Cancers |
| Duodenal Cancer | 32 | 1.0% | Cancers |
| Cervix carcinoma | 28 | 0.9% | Cancers |
| Carotid Artery Diseases | 27 | 0.9% | Heart Diseases |
| Glycogen Storage Disease Type VI | 27 | 0.9% | Liver Diseases |
| Breast Carcinoma | 27 | 0.9% | Cancers |
| Congestive heart failure | 26 | 0.8% | Heart Diseases |
| Endometrial Carcinoma | 26 | 0.8% | Cancers |
| Chronic Kidney Diseases | 25 | 0.8% | Kidney Diseases |
| Anal squamous cell carcinoma | 24 | 0.8% | Cancers |
| Malignant mesothelioma | 24 | 0.8% | Cancers |
| Malignant Pleural Mesothelioma | 24 | 0.8% | Cancers |
| Squamous cell carcinoma | 24 | 0.8% | Cancers |
| Squamous cell carcinoma - category | 24 | 0.8% | Cancers |
| Carcinoma of Male Breast | 23 | 0.7% | Cancers |
| Childhood Breast Carcinoma | 23 | 0.7% | Cancers |
| End Stage Liver Disease | 21 | 0.7% | Liver Diseases |
| Nasopharyngeal carcinoma | 21 | 0.7% | Cancers |
| Carcinoma of anal canal | 19 | 0.6% | Cancers |
| Infant Acute Lymphoblastic Leukemia | 18 | 0.6% | Cancers |
| Carcinoma of soft palate | 18 | 0.6% | Cancers |
| Carcinoma of lung | 17 | 0.5% | Cancers |
| Metastatic Prostate Carcinoma | 17 | 0.5% | Cancers |
| Pancreatic carcinoma | 17 | 0.5% | Cancers |
| Prostate carcinoma | 17 | 0.5% | Cancers |
| Colon Carcinoma | 16 | 0.5% | Cancers |
| Colorectal Carcinoma | 16 | 0.5% | Cancers |
| Urethral Carcinoma | 14 | 0.5% | Cancers |
| Carcinoma of subglottis | 13 | 0.4% | Cancers |
| Conventional (Clear Cell) Renal Cell Carcinoma | 13 | 0.4% | Cancers |
| Experimental Organism Basal Cell Carcinoma | 13 | 0.4% | Cancers |
| Maxillary Sinus Carcinoma | 13 | 0.4% | Cancers |
| Thyroid carcinoma | 13 | 0.4% | Cancers |
| Bronchioloalveolar Adenocarcinoma | 12 | 0.4% | Cancers |
| CARCINOMA OF VULVA | 12 | 0.4% | Cancers |
| Prediabetes syndrome | 12 | 0.4% | Diabetes |
| Carcinoma of bladder | 11 | 0.4% | Cancers |
| Malignant melanoma of anus | 11 | 0.4% | Cancers |
| melanoma | 11 | 0.4% | Cancers |
| Renal Pelvis Carcinoma | 11 | 0.4% | Cancers |
| Adrenal Cancer | 10 | 0.3% | Cancers |
| Diabetes with coma (disorder) | 10 | 0.3% | Diabetes |
| Nasal cavity carcinoma | 10 | 0.3% | Cancers |
| Retroperitoneal Carcinoma | 10 | 0.3% | Cancers |
| Skin Basal Cell Carcinoma | 10 | 0.3% | Cancers |
| Acute Kidney Insufficiency | 9 | 0.3% | Kidney Diseases |
| Non-Small Cell Lung Carcinoma | 9 | 0.3% | Cancers |
| Cancer Other | 8 | 0.3% | Cancers |
| Childhood Bladder Carcinoma | 8 | 0.3% | Cancers |
| Gestational Diabetes | 8 | 0.3% | Diabetes |
| Pre-Gestational Diabetes | 8 | 0.3% | Diabetes |
| Basal cell carcinoma | 7 | 0.2% | Cancers |
| Carcinoma in Situ | 7 | 0.2% | Cancers |
| Cystic Kidney Diseases | 7 | 0.2% | Kidney Diseases |
| Endometrial Endometrioid Adenocarcinoma | 7 | 0.2% | Cancers |
| Kidney Diseases | 7 | 0.2% | Kidney Diseases |
| Medullary carcinoma of thyroid | 7 | 0.2% | Cancers |
| Papillary Serous Cystadenocarcinoma | 7 | 0.2% | Cancers |
| Parathyroid Gland Adenocarcinoma | 7 | 0.2% | Cancers |
| Polycystic liver disease | 7 | 0.2% | Liver Diseases |
| Renal cell carcinoma 1 | 7 | 0.2% | Cancers |
| Squamous Cell Carcinoma of Unknown Primary | 7 | 0.2% | Cancers |
| Uterine Cancer | 7 | 0.2% | Cancers |
| Canine Transitional Cell Carcinoma | 6 | 0.2% | Cancers |
| Chronic Kidney Insufficiency | 6 | 0.2% | Kidney Diseases |
| Gastroesophageal Junction Adenocarcinoma | 6 | 0.2% | Cancers |
| Heart Diseases | 6 | 0.2% | Heart Diseases |
| Hypopharyngeal Cancer | 6 | 0.2% | Cancers |
| Hypopharyngeal Carcinoma | 6 | 0.2% | Cancers |
| Left-Sided Heart Failure | 6 | 0.2% | Heart Diseases |
| Lentigo maligna melanoma | 6 | 0.2% | Cancers |
| Malignant Kidney Mixed Epithelial and Stromal Tumor | 6 | 0.2% | Cancers |
| Pituitary carcinoma | 6 | 0.2% | Cancers |
| Primary Prostate Urothelial Carcinoma | 6 | 0.2% | Cancers |
| Stage I Lung Non-Small Cell Cancer AJCC v7 | 6 | 0.2% | Cancers |
| Transitional cell carcinoma of kidney | 6 | 0.2% | Cancers |
| Transitional cell carcinoma of ureter | 6 | 0.2% | Cancers |
| bilateral breast cancer | 5 | 0.2% | Cancers |
| Carcinoma of supraglottis | 5 | 0.2% | Cancers |
| Endocervical Carcinoma | 5 | 0.2% | Cancers |
| Heart valve disease | 5 | 0.2% | Heart Diseases |
| Invasive Ductal Breast Carcinoma | 5 | 0.2% | Cancers |
| Liver Cirrhosis | 5 | 0.2% | Liver Diseases |
| Non-small cell carcinoma | 5 | 0.2% | Cancers |
| Paranasal Sinus Cancer | 5 | 0.2% | Cancers |
| Polycystic Kidney Diseases | 5 | 0.2% | Kidney Diseases |
| Squamous cell carcinoma of buccal mucosa | 5 | 0.2% | Cancers |
| Stage I Uterine Corpus Cancer AJCC v6 | 5 | 0.2% | Cancers |
| Adenocarcinoma of prostate | 4 | 0.1% | Cancers |
| Carcinoma in situ of penis | 4 | 0.1% | Cancers |
| Cervical Adenocarcinoma In Situ | 4 | 0.1% | Cancers |
| Dipsogenic Diabetes Insipidus | 4 | 0.1% | Diabetes |
| Endometrioid carcinoma of prostate | 4 | 0.1% | Cancers |
| Liver diseases | 4 | 0.1% | Liver Diseases |
| Luminal A Breast Carcinoma | 4 | 0.1% | Cancers |
| Luminal B Breast Carcinoma | 4 | 0.1% | Cancers |
| Lung Hilum Carcinoma | 4 | 0.1% | Cancers |
| Metastatic Colon Adenocarcinoma | 4 | 0.1% | Cancers |
| Metastatic Gastric Adenocarcinoma | 4 | 0.1% | Cancers |
| Metastatic Lung Carcinoma | 4 | 0.1% | Cancers |
| Metastatic Prostate Adenocarcinoma | 4 | 0.1% | Cancers |
| Metastatic Rectal Adenocarcinoma | 4 | 0.1% | Cancers |
| Noninfiltrating Intraductal Carcinoma | 4 | 0.1% | Cancers |
| Papillary carcinoma in situ | 4 | 0.1% | Cancers |
| Papillary carcinoma of the breast | 4 | 0.1% | Cancers |
| Papillary Carcinoma of the Penis | 4 | 0.1% | Cancers |
| Parotid Gland Carcinoma | 4 | 0.1% | Cancers |
| Prostate Acinar Adenocarcinoma | 4 | 0.1% | Cancers |
| Stage 0 Colon Cancer AJCC v6 and v7 | 4 | 0.1% | Cancers |
| Stage 0 Lip Carcinoma | 4 | 0.1% | Cancers |
| Stage 0 Pharynx Carcinoma | 4 | 0.1% | Cancers |
| Stage I Pharyngeal Cancer | 4 | 0.1% | Cancers |
| Stage I Prostate Adenocarcinoma AJCC v7 | 4 | 0.1% | Cancers |
| Stage II Lung Non-Small Cell Cancer AJCC v7 | 4 | 0.1% | Cancers |
| Stage II Pharynx Carcinoma | 4 | 0.1% | Cancers |
| Adenoid Cystic Breast Carcinoma | 3 | 0.1% | Cancers |
| Bladder Urachal Carcinoma | 3 | 0.1% | Cancers |
| Breast Adenocarcinoma | 3 | 0.1% | Cancers |
| Cancer of Nasopharynx | 3 | 0.1% | Cancers |
| Chronic liver disease | 3 | 0.1% | Liver Diseases |
| Endometrial Serous Adenocarcinoma | 3 | 0.1% | Cancers |
| Gallbladder Carcinoma | 3 | 0.1% | Cancers |
| Genitourinary Cancer | 3 | 0.1% | Cancers |
| Infratentorial Cancer | 3 | 0.1% | Cancers |
| Invasive Carcinoma | 3 | 0.1% | Cancers |
| Invasive carcinoma of breast | 3 | 0.1% | Cancers |
| Invasive Prostate Carcinoma | 3 | 0.1% | Cancers |
| Labia Minora Carcinoma | 3 | 0.1% | Cancers |
| Papillary renal cell carcinoma type 2 | 3 | 0.1% | Cancers |
| Precancerous Polyp | 3 | 0.1% | Cancers |
| Stage I Carcinoma of Bladder | 3 | 0.1% | Cancers |
| Stage I Nasopharyngeal Carcinoma | 3 | 0.1% | Cancers |
| Stage II Bladder Cancer AJCC v6 and v7 | 3 | 0.1% | Cancers |
| Stage IV Thyroid Gland Papillary Carcinoma AJCC v7 | 3 | 0.1% | Cancers |
| Type 1 Papillary Renal Cell Carcinoma | 3 | 0.1% | Cancers |
| Acute Coronary Syndrome | 2 | 0.1% | Heart Diseases |
| Breast Carcinoma Metastatic in the Bone | 2 | 0.1% | Cancers |
| Breast Carcinoma Metastatic in the Liver | 2 | 0.1% | Cancers |
| Breast Carcinoma Metastatic in the Spine | 2 | 0.1% | Cancers |
| Carcinoma breast stage IV | 2 | 0.1% | Cancers |
| Carcinoma in situ of uterine cervix | 2 | 0.1% | Cancers |
| Carcinoma of glottis | 2 | 0.1% | Cancers |
| Carcinoma of unknown primary | 2 | 0.1% | Cancers |
| Carcinomatosis | 2 | 0.1% | Cancers |
| Cervical Mucinous Adenocarcinoma | 2 | 0.1% | Cancers |
| Cervical Serous Adenocarcinoma | 2 | 0.1% | Cancers |
| Childhood Esophageal Carcinoma | 2 | 0.1% | Cancers |
| Chronic kidney disease stage 1 | 2 | 0.1% | Kidney Diseases |
| Chronic kidney disease stage 2 | 2 | 0.1% | Kidney Diseases |
| Coronary Stenosis | 2 | 0.1% | Heart Diseases |
| Endometrial adenocarcinoma | 2 | 0.1% | Cancers |
| Esophageal carcinoma | 2 | 0.1% | Cancers |
| Feline Bronchioloalveolar Lung Carcinoma | 2 | 0.1% | Cancers |
| Grade 1 Colon Adenocarcinoma | 2 | 0.1% | Cancers |
| Grade 1 Rectal Adenocarcinoma | 2 | 0.1% | Cancers |
| Grade 2 Colon Adenocarcinoma | 2 | 0.1% | Cancers |
| Grade 2 Colorectal Adenocarcinoma | 2 | 0.1% | Cancers |
| Grade 2 Rectal Adenocarcinoma | 2 | 0.1% | Cancers |
| Infiltrating papillary adenocarcinoma | 2 | 0.1% | Cancers |
| Intraepithelial Squamous Cell Carcinoma | 2 | 0.1% | Cancers |
| Invasive Lobular Breast Carcinoma | 2 | 0.1% | Cancers |
| Liver Failure | 2 | 0.1% | Liver Diseases |
| Low Grade Ductal Breast Carcinoma In Situ | 2 | 0.1% | Cancers |
| Lung Spindle Cell Carcinoma | 2 | 0.1% | Cancers |
| Malignant melanoma of vulva | 2 | 0.1% | Cancers |
| Meningeal Carcinomatosis | 2 | 0.1% | Cancers |
| Metastatic Endometrial Carcinoma | 2 | 0.1% | Cancers |
| Moderately Differentiated Prostate Adenocarcinoma | 2 | 0.1% | Cancers |
| Nonocclusive coronary artery atherosclerosis | 2 | 0.1% | Heart Diseases |
| Non-Small Cell Adenocarcinoma | 2 | 0.1% | Cancers |
| Pancreatic Moderately Differentiated Ductal Adenocarcinoma | 2 | 0.1% | Cancers |
| Pleural Carcinomatosis | 2 | 0.1% | Cancers |
| Polycystic kidney disease | 2 | 0.1% | Kidney Diseases |
| POLYCYSTIC KIDNEY DISEASE 1 | 2 | 0.1% | Kidney Diseases |
| POLYCYSTIC KIDNEY DISEASE 4 | 2 | 0.1% | Kidney Diseases |
| POLYCYSTIC KIDNEY DISEASE 5 | 2 | 0.1% | Kidney Diseases |
| Premature coronary artery atherosclerosis | 2 | 0.1% | Heart Diseases |
| Rectal Sarcomatoid Carcinoma | 2 | 0.1% | Cancers |
| Rectosigmoid Carcinoma | 2 | 0.1% | Cancers |
| Recurrent Vulvar Carcinoma | 2 | 0.1% | Cancers |
| Sarcomatoid Carcinoma of the Penis | 2 | 0.1% | Cancers |
| Small cell carcinoma of lung | 2 | 0.1% | Cancers |
| Small intestine carcinoma | 2 | 0.1% | Cancers |
| Stage 0 Bladder Urothelial Carcinoma AJCC v6 and v7 | 2 | 0.1% | Cancers |
| Stage 0 Non-Small Cell Lung Cancer AJCC v6 and v7 | 2 | 0.1% | Cancers |
| Stage I Bladder Urothelial Carcinoma AJCC v6 and v7 | 2 | 0.1% | Cancers |
| Stage I Esophageal Cancer | 2 | 0.1% | Cancers |
| Stage I Lung Small Cell Carcinoma AJCC v7 | 2 | 0.1% | Cancers |
| Stage IA Uterine Corpus Cancer AJCC v7 | 2 | 0.1% | Cancers |
| Stage II Bladder Urothelial Carcinoma AJCC v6 and v7 | 2 | 0.1% | Cancers |
| Stage IV Bladder Urothelial Carcinoma AJCC v7 | 2 | 0.1% | Cancers |
| Thymic Sarcomatoid Carcinoma | 2 | 0.1% | Cancers |
| Transitional cell carcinoma of bladder | 2 | 0.1% | Cancers |
| Type II Endometrial Adenocarcinoma | 2 | 0.1% | Cancers |
| Undifferentiated Carcinoma of Colon | 2 | 0.1% | Cancers |
| Well Differentiated Prostate Adenocarcinoma | 2 | 0.1% | Cancers |
| Acute lymphocytic leukemia | 1 | 0.0% | Cancers |
| Adult Hepatocellular Carcinoma | 1 | 0.0% | Cancers |
| Bilateral Carcinoma | 1 | 0.0% | Cancers |
| Buccal Mucosa Verrucous Carcinoma | 1 | 0.0% | Cancers |
| Canine Hepatocellular Carcinoma | 1 | 0.0% | Cancers |
| Cervical Clear Cell Adenocarcinoma | 1 | 0.0% | Cancers |
| Childhood Hepatocellular Carcinoma | 1 | 0.0% | Cancers |
| Contralateral Breast Carcinoma | 1 | 0.0% | Cancers |
| Ductal Breast Carcinoma | 1 | 0.0% | Cancers |
| Ductal Carcinoma In Situ with Microinvasion | 1 | 0.0% | Cancers |
| Exocervical Carcinoma | 1 | 0.0% | Cancers |
| Extrapulmonary Small Cell Carcinoma | 1 | 0.0% | Cancers |
| Labia Majora Carcinoma | 1 | 0.0% | Cancers |
| Liver carcinoma | 1 | 0.0% | Cancers |
| Metastatic Anal Canal Carcinoma | 1 | 0.0% | Cancers |
| Metastatic Carcinoma | 1 | 0.0% | Cancers |
| Metastatic Colon Carcinoma | 1 | 0.0% | Cancers |
| Metastatic Lip Carcinoma | 1 | 0.0% | Cancers |
| Metastatic Penile Cancer | 1 | 0.0% | Cancers |
| Sebaceous Adenocarcinoma | 1 | 0.0% | Cancers |
| Thyroid Gland Adenocarcinoma | 1 | 0.0% | Cancers |
| Unilateral Breast Carcinoma | 1 | 0.0% | Cancers |
| Verrucous carcinoma | 1 | 0.0% | Cancers |

# Appendix B

Detailed results for each model tested in this case study are included here. They are organized by model group (multiclass or binary) then by model.

## Multiclass Classification Results

**Table B1**   
Multiclass Logistic Regression Classification Report

Table

Description automatically generated

**Figure B1**  
Multiclass Logistic Regression Confusion Matrices

Calendar

Description automatically generated with medium confidence Chart

Description automatically generated

**Table B2**   
Multiclass Naïve Bayes Classification Report

Table

Description automatically generated

**Figure B2**Multiclass Naïve Bayes Confusion Matrices

A picture containing calendar

Description automatically generated Chart, bar chart

Description automatically generated

**Table B3**   
Multiclass Decision Tree Classification Report

Table

Description automatically generated

**Figure B3**  
Multiclass Decision Tree Confusion Matrices

A picture containing scatter chart

Description automatically generated Chart, bar chart

Description automatically generated

**Table B4** Multiclass Neural Networks Training Classification Report

Table

Description automatically generated

**Figure B4**  
Multiclass Neural Networks Training Confusion Matrices

Calendar

Description automatically generated Chart

Description automatically generated

**Table B5**Multiclass Neural Networks Testing Classification Report

Table

Description automatically generated

**Figure B5**Multiclass Neural Networks Testing Confusion Matrices

Calendar

Description automatically generated Chart, bar chart

Description automatically generated

## Binary Classification Results

**Table B6**Binary Logistic Regression Classification Report

Table

Description automatically generated

**Figure B6**Binary Logistic Regression Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

**Table B7**Binary Naive Bayes Classification Report

Table

Description automatically generated

**Figure B7**  
Binary Naïve Bayes Regression Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

**Table B8**Binary Decision Tree Classification Report

Table

Description automatically generated

**Figure B8**  
Binary Decision Trees Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

**Table B9**  
Binary Neural Networks Training Classification Report

Table

Description automatically generated

**Figure B9**Binary Neural Networks Training Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

**Table B10**  
Binary Neural Networks Testing Classification Report

Table

Description automatically generated

**Figure B10**Binary Neural Networks Testing Confusion Matrices

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

# Appendix C

Code and data files for this paper can be found on GitHub using the following link:

<https://github.com/koreybernhardt/capstone>. Individual files are listed below and also include a direct link.

* [DS785\_Final\_Code.ipynb](https://github.com/koreybernhardt/capstone/blob/main/DS785_Final_Code.ipynb): Jupyter notebook with all code used in analysis.
* [disease\_aggregation.csv](https://github.com/koreybernhardt/capstone/blob/main/disease_aggregation.csv): Mapping of specific diseases to one of five high mortality diseases.
* [disease\_counts.csv](https://github.com/koreybernhardt/capstone/blob/main/disease_counts.csv): All diseases assigned to transcriptions with count of transcriptions.
* [mtsamples.csv](https://github.com/koreybernhardt/capstone/blob/main/mtsamples.csv): Source data obtained from Kaggle.