August 2020

Machine-Learning-based Prediction of Sepsis Events from Vertical Clinical Trial Data: a Naïve Approach

Tyler Michael Gaddis
University of Wisconsin-Milwaukee

MACHINE-LEARNING-BASED SEPSIS PREDICTION USING VERTICAL CLINICAL TRIAL DATA: A NAÏVE APPROACH

by

Tyler Gaddis

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

in Health Care Informatics

at

The University of Wisconsin – Milwaukee

August 2020

ABSTRACT

MACHINE-LEARNING-BASED SEPSIS PREDICTION USING VERTICAL CLINICAL TRIAL DATA: A NAÏVE APPROACH

by

Tyler Gaddis

The University of Wisconsin – Milwaukee, 2020
Under the Supervision of Jake Luo, PhD

Sepsis is a potentially life-threatening condition characterized by a dysregulated, disproportionate immune response to infection by which the afflicted body attacks its own tissues, sometimes to the point of organ failure, and in the worst cases, death. According to the Centers for Disease Control and Prevention (CDC) Sepsis is reported to kill upwards of 270,000 Americans annually, though this figure may be greater given certain ambiguities in the current accepted diagnostic framework of the disease.

This study attempted to first establish an understanding of past definitions of sepsis, and to then recommend use of machine learning as integral in an eventual amended disease definition. Longitudinal clinical trial data (n_{trials}=30,915) were vectorized into a machine-readable format compatible with predictive modeling, selected and reduced in dimension, and used to predict incidences of sepsis via application of several machine learning models: logistic regression, support vector machines (SVM), naïve Bayes Classifier, decision trees, and random forests. The intent of the study was to identify possible predictive features for sepsis via comparative analysis of different machine learning models, and to recommend subsequent study of sepsis prediction using the training model on new data (non-clinical-trial-derived) in

the same format. If the models can be generalized to new data, it stands to assume they could eventually become clinically useful. In referencing F1 scores and recall scores, the random forest classifier was the best performer among this cohort of models.

To my wife, Drew: You are strong, and good, and true, and the reason I am who I've become. I love you.

To Mom and Dad: I revel at your support, patience, love, forgiveness, and generosity. I can only hope I do the same.

To Sarah and Roger: Thank you for your unequivocal belief in me.

Deyr fé, deyja frændr, deyr sjalfr it sama,
en orðstirr deyr aldregi, hveim er sér góðan getr.

Deyr fé, deyja frændr, deyr sjalfr it sama,
ek veit einn, at aldrei deyr: dómr um dauðan hvern.

TABLE OF CONTENTS

ABSTRACT ii

DEDICATION v

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF ABBREVIATIONS xi

CHAPTER

1. INTRODUCTION 1

1.1. Relevant Sepsis Statistics 2

1.2. Sepsis: A Formal Definition 3

1.2.1. Sepsis-1 3

1.2.2. Sepsis-2 5

1.2.3. Sepsis-3 7

1.3. A Critique of the 2016 Sepsis Definition 9

2. METHODS 12

2.1. Data Source 12

2.2. Data Transformation and Preprocessing 13

2.3. Feature Selection 18

2.3.1.Variance-Based Feature Selection 19

2.3.2.Correlation-Based Feature Selection 19

2.4. Dimensionality Reduction 21

2.4.1.Truncated SVD 21

2.5. Cross-Validation 24
2.6. Models 25
2.6.1. Logistic Regression 26
2.6.1.1. Regularization 28
2.6.2. Support Vector Machine (SVM) 30
2.6.2.1. Linear Kernel 32
2.6.2.2. Radial Basis Function Kernel 33
2.6.3. Naïve Bayes 33
2.6.4. Decision Tree 34
2.6.5. Random Forest 36
3. OBJECTIVES 37
4. RESULTS 38
4.1. Logistic Regression Sag Solver 39
4.2. Support Vector Machine: Linear Kernel 41
4.3. Naïve Bayes: Gaussian Classifier 43
4.4. Decision Tree: Gini Impurity 45
4.5. Random Forest: Gini Impurity 47
5. DISCUSSION 49
6. CONCLUSION 51

REFERENCES 53

LIST OF FIGURES

Figure 1. SIRS, Sepsis, and Infection Venn Diagram 4

Figure 2. Raw Sparse Data 14

Figure 3. Feature Headers 14

Figure 4. Preparing the Data 15-17

Figure 5. Pre-Feature-Selected Data 17

Figure 6. Simplified linear algebra of the Truncated SVD algorithm 22

Figure 7. Simplified Truncated SVD Component Vector Addition 23

Figure 8. K-Fold Cross Validation 24

Figure 9. Sigmoid Function 26

Figure 10. Kernel Trick for Hyperplane Identification 31

Figure 11. Various SVM Kernel Applications on Iris Dataset 32

Figure 12. Decision tree: Unpruned 35

Figure 13. AUCROC For Logistic Regression Sag Solver 39

Figure 14. Logistic Regression Sag Solver Confusion Matrix 39

Figure 15. Normalized Confusion Matrix for Sag Solver 40

Figure 16. AUCROC for SVM.SVC Linear Kernel 41

Figure 17. SVM.SVC Linear Kernel Confusion Matrix 41

Figure 18. Normalized SVM.SVC Linear Kernel Confusion Matrix 42

Figure 19. AUCROC For Naïve Bayes Gaussian Classifier 43

Figure 20. Naïve Bayes Gaussian Classifier Confusion Matrix 43

Figure 21. Normalized Naïve Bayes Gaussian Classifier Confusion Matrix 44

Figure 22. AUCROC for Gini Impurity Depth 10 Decision Tree 45

Figure 23. Gini Impurity Depth 10 Decision Tree Confusion Matrix..... 45

Figure 24. Normalized Gini Impurity Depth 10 Decision Tree Confusion Matrix 46

Figure 25. AUROC for Random Forest Gini Impurity 47

Figure 26. Random Forest Gini Impurity Confusion Matrix..... 47

Figure 27. Normalized Random Forest Gini Impurity Confusion Matrix 48

LIST OF TABLES

Table 1. SIRS Criteria 4

Table 2. Sepsis-2 Criteria 6

Table 3. Features and F-Statistics 20

Table 4. Optimized Model Results Table 38

LIST OF ABBREVIATIONS

LR	Logistic Regression
SVM	Support Vector Machine
SVD	Singular Value Decomposition
RBF	Radial Basis Function Kernel
NB	Naïve Bayes
DT	Decision Tree
RF	Random Forest
sklearn	Scikit-Learn Machine Learning Library
XGB	XGBoost = Extreme Gradient Boosting/Boosted Trees
ACCP-SCCM	American College of Chest Physicians-Society of Critical Care Medicine
SIRS	Systemic Inflammatory Response Syndrome
CDC	Centers for Disease Control and Prevention
JAMA	The Journal of the American Medical Association
SOFA	Sepsis-related Organ Failure Assessment
qSOFA	quick Sepsis-related Organ Failure Assessment

1. INTRODUCTION

Sepsis is a clinical syndrome of exaggerated and life-threatening systemic immune responses launched by the body against its own tissues on encountering an infection ultimately resulting in organ damage, organ failure, or death. It is a syndrome comprising myriad combinations of clinical symptoms in patients suffering from infection, rendering its precise pathophysiologic definition and subsequent treatment elusive and tenuous at best (Singer et al., 2016). No single system, pathogen, mediator, or pathway have been isolated as preeminent drivers of sepsis pathophysiology (Remick, 2007). Sepsis diagnosis is made ambiguous due to its shared symptoms with other comorbidity pathophysiologies, and modern pervasive use of antibiotics producing false negative culture results (Vincent, 2016).

The urgency of sepsis and a valid sepsis diagnostic tool is underscored by several facts: that all of the body's organ systems are susceptible to it; that the only requirement for sepsis onset is bacterial, viral, fungal, or parasitic infection; that its frequency is increasing due to an aging population long benefitting from chronic condition management (here it is hypothesized that the conditions being managed may be predictive factors for sepsis); that until a 2016 task force that redefined criteria for sepsis (Sepsis-3), the syndrome definition was over-reliant on inflammation as its baseline assessment criterion, and on a misguided spectrum model of disease (Singer et al., 2016); that up to and beyond this task force the definition lacked general consensus pertaining to its usefulness for clinical diagnosis versus prognosis; that a demonstrable heterogeneity of inflammatory response and cellular changes in organ tissues of septic patients exists (Remick, 2007), and thus, that sepsis can initially be clinically indistinguishable from systemic inflammation from non-infectious causes (Lopansri, Miller, &

Brandon, 2019); that survivors of sepsis remain susceptible to subsequent chronic physiological, psychological and cognitive ailments.

Because sepsis represents a relatively common but acute and often lethal clinical syndrome, continued efforts must be leveraged to identify its precise etiology and pathophysiology. As a contribution to this effort, this case study attempts to identify predictors for sepsis from a large clinical trial dataset (n = 30,915) in the form of severe and less-severe adverse events, trial stage, preexisting conditions, and interventions used for trial health outcomes. If comorbidities of sepsis are isolated in a large enough sample size, it could be argued that sepsis ought to be treated according to site-specific biomarkers of, conditions of, and/or proximity to organ systems where the syndrome develops.

1.1. RELEVANT SEPSIS STATISTICS

Sepsis represents a significant tax on the American healthcare infrastructure and deserves a corresponding magnitude of attention. In 2013 alone, sepsis accounted for $24 billion of total US hospital costs (Paoli et al., 2018). One two-cohort study (Kaiser Permanente Northern California, n=482,828; Healthcare Cost and Utilization Project Nationwide Inpatient Sample, n=6,500,000) found that upwards of 50% of all hospital deaths are attributable to sepsis (Liu et al., 2014).

According to the CDC (Centers for Disease Control and Prevention), 1.7 million American adults develop sepsis annually, 270,000 of whom die from the disease, amounting to a mortality rate of nearly 16% (Centers for Disease Control and Prevention, 2020). Sepsis is most commonly seen in adults aged 65 or older, immunocompromised and chronically ill patients,

and in infants. Signs and symptoms include tachycardia, disorientation/confusion, discomfort, fever or hypothermia, dyspnea, and perspiration.

1.2. SEPSIS: A FORMAL DEFINITION

Sepsis has been in the medical consciousness for millennia and has been formally defined several times. Louis Pasteur's germ theory of disease attributed infection to harmful microbes, thus spurring the first science-derived, empirically driven pursuit of sepsis comprehension. The first modern definition of sepsis, posited by Hugo Schottmüller in 1914, was than more modern definitions:

> *Sepsis is present if a focus has developed from which pathogenic bacteria, constantly or periodically, invade the blood stream in such a way that this causes subjective and objective symptoms.*
>
> (Gyawali et al., 2019)

More specifically, its definition has evolved as the pathophysiology of the syndrome and pathobiology of affected tissues have enjoyed greater understanding. Subsequent official definitions followed Schottmüller's, most notably Sepsis-1, Sepsis-2, and Sepsis-3.

1.2.1. Sepsis-1

In 1991 the first consensus definition of sepsis was established at an American College of Chest Physicians-Society of Critical Care Medicine (ACCP-SCCM) conference helmed by Roger Bone. Bone made an argument for the improved definition of sepsis, and the importance of precision of language in defining it (Bone, 1991). According to the proposed definition, sepsis was a spectrum of systemic responses to infection ranging from systemic inflammatory response syndrome (SIRS), to severe sepsis (sepsis complicated by organ dysfunction), to septic shock ("sepsis-induced hypotension persisting despite adequate fluid resuscitation") (Bone et

al., 1992). SIRS was the first of these stages of sepsis and so a definition considering the host's inflammatory response to infection as its foundational attribute followed. Sepsis-1 was ideally framed to treat sepsis and SIRS as non-disjoint systemic responses to environmental factors including but not limited to infection (Bone et al., 1992).

Table 1. SIRS Criteria

Sepsis-1
Sepsis is a systemic inflammatory response in the presence of infection
SIRS criteria
Temperature > 38°C or < 36°C
Heart rate > 90/minute
Respiratory rate > 20/minute (or $PaCO_2$ < 32 mmHg)
WBC > 12,000/μL or < 4,000/μL (or > 10% immature bands)

(Carneiro, Povoa, & Gomes, 2017)

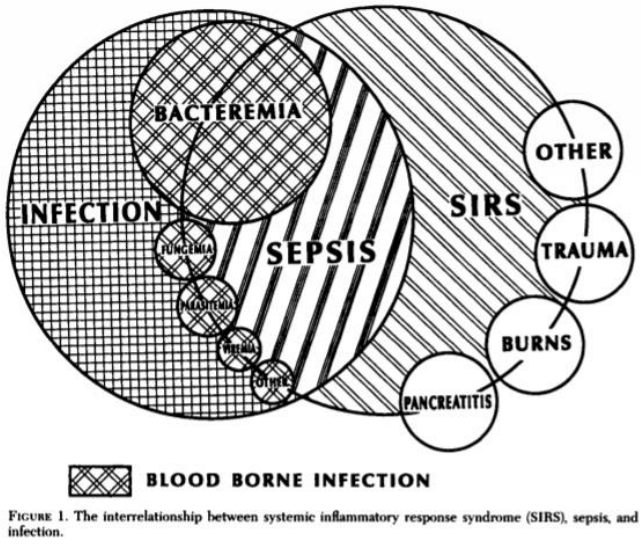


FIGURE 1. The interrelationship between systemic inflammatory response syndrome (SIRS), sepsis, and infection.

Figure 1. SIRS, Sepsis, and Infection Venn Diagram

(Bone et al., 1992)

Four SIRS criteria were established: tachycardia (resting heart rate of over 90 beats per minute), tachypnea (respiration rate of over 20 times per minute), body core temperature dysregulation (fever: core temperature above 38°C; hypothermia: core temperature below

36°C), and leukocytosis, leukopenia, or bandemia (leukocytes in greater concentrations than 1200/mm^3, leukocytes in lower concentrations than 4000/mm^3, or bandemia of ≥10%m, respectively). The 1991 convention asserts that to be clinically diagnosed as SIRS-positive, patients must exhibit two or more of these criteria (Marik and Taeb, 2017).

But the SIRS criteria proved to be far too general. If any two of the four SIRS criteria were observed in patients, SIRS criteria were theoretically met, and thus the first "stage" of sepsis would have been recorded as observed in such patients. For example, on average, up to 90% of intensive care unit (ICU) patients exhibit symptoms meeting SIRS criteria, to the extent that they are eligible for sepsis diagnosis according to the Sepsis-1-SIRS criteria (Sprung et al., 2006). Given that most infections induce some form of SIRS criteria (for example, tachycardia, fever, high leukocyte counts) yet such infections less often result in actual recorded incidence of sepsis (in the most relevant definition's sense of the word), Sepsis-1 and its heavy reliance on SIRS criteria represents a mischaracterization of infection and inflammation as sepsis. A new definition was need.

1.2.2. Sepsis-2

In 2001, a task force met to address the limitations of the Sepsis-1 definition, and in so doing attempted a reformed definition of sepsis, later coined "Sepsis-2" by the Journal of the American Medical Association in 2016. The task force at the 2001 consensus conference sought a more precise definition of sepsis via a thorough consideration of all clinical factors associated with it. But what was intended as a conference to reform the definition of sepsis as a condition commensurate with a few comorbidities resulted in loss of specificity and clarity as to what constituted the condition. The Sepsis-2 baseline criterium was infection, included the four SIRS

criteria from Sepsis-1, and several coinciding SIRS-related symptoms. These other symptoms were grouped by “general parameters”, “inflammatory parameters”, “hemodynamic parameters”, “tissue perfusion parameters”, and “organ dysfunction parameters”, and are summarized in the following table:

Table 2. Sepsis-2 Criteria

Sepsis-2	
General signs and symptoms	Hemodynamic variables
Fever (central temperature > 38.3°C)	Arterial hypotension (systolic < 90 mmHg, MAP < 70 mmHg, or systolic reduction > 40 mmHg in adults or < 2 SD of the normal value for age)
Hypothermia (central temperature < 36°C)	SvO_2 < 70%
Heart rate > 90/minute or > 2 SD above the normal value for age	Cardiac index > 3.5 L/min/m²
Tachypnea	Indicators of organ dysfunction
Edema or positive fluid balance (> 20 mL/kg 24 hours)	Arterial hypoxemia (PaO_2/FiO_2 < 300)
Hyperglycemia (glycemia > 120 mg/dL) in the absence of diabetes	Abnormal state of consciousness
Inflammation markers	Acute oliguria (urine output < 0.5 mL/kg/hour)
Leukocytosis (> 12,000/μL) or leukopenia (< 4,000/μL)	Elevated creatinine > 0.5 mg/dL
Normal leukocytes but > 10% immature bands	Coagulation disorders (INR > 1.5/aPTT > 60 s)
Serum C-reactive protein > 2 SD above the normal value	Thrombocytopenia (< 100,000/μL)
Plasma procalcitonin > 2 SD above the normal value	Hyperbilirubinemia (> 4 mg/dL or 70 mmol/L)
	Indicators of tissue perfusion
	Hyperlactatemia (> 1 mmol/L)
	Reduced capillary refill and mottled skin

(Carneiro, Povoa, & Gomes, 2017)

Though intended as a comprehensive reference for indicators of sepsis, many of the criteria from this list were consistent with normal immune responses to infection. Despite its intention for specificity, it broadened the scope of symptoms and subsequently was in danger of identifying infection paired with any of the listed comorbidities as proxies for sepsis. To that end, the conference was considered a failure, and Sepsis-1 persisted as the most relevant definition for sepsis (Vincent, et al., 2013). Between 2001 and 2016 (Sepsis-3), advances in

understanding of sepsis pathophysiology, etiology, pathobiology, and immunology were made, which merited a new definition for the condition.

1.2.3. Sepsis-3

In 2016 the Journal of the American Medical Association (JAMA) proposed a third definition of Sepsis that abandoned an argued overreliance on SIRS/inflammation, and instead opted to treat the syndrome not as a continuum, but as a "life-threatening organ dysfunction caused by a dysregulated host response to infection" (Singer et al., 2016). More specifically this definition attempts to discriminate between past notions of sepsis and non-sepsis-related infection, to account for pro- and anti-inflammatory responses associated with sepsis (which otherwise would have been confounding, given previous definition's reliance on inflammation/SIRS criteria), and to establish the "primacy of the nonhomeostatic host response to infection, the potential lethality that is considerably in excess of a straightforward infection, and the need for urgent recognition" (Singer et al., 2016). This definition is argued to be the most and accurate and practical addendum to AMA-sanctioned formal sepsis definitions because it accounts for modern conventional wisdom while yielding that less understood influences (e.g. genetic or cellular influences) could yet impact sepsis pathophysiology and pathobiology. In addition, the definition was designed to address severe variations in sepsis incidence and mortality attributed to a lack of standardized definitions for sepsis and septic shock. (Singer et al., 2016)

By restructuring and reforming the existing sepsis framework to instead focus on infection, host response, and organ dysfunction, JAMA cited an improved understanding of pathobiology

("organ function, morphology, cell biology, biochemistry, immunology and circulation") as the chief impetus for its revision (Singer et al., 2016).

Because this definition rejected the limited emphasis on inflammation placed by the Sepsis-1 and Sepsis-2 definitions, new criteria constituting sepsis and septic shock were required.

Given that sepsis phenotypes differ across patient populations manifesting a range of different comorbidities, interventions, and infections, a broader understanding of sepsis was pursued. In response to this need, the 2016 task force suggested largely abandoning SIRS criteria in favor of the Sepsis-related Organ Failure Assessment (SrOFA), renamed the Sequential Organ Failure Assessment (SOFA) as primary criteria for diagnosis. SOFA criteria were designed to identify signs of all previously identified sepsis symptoms, namely "infection, host response, and organ dysfunction". Under these guidelines, if a patient presents with a SOFA score greater than 2, they are immediately assigned a 10% mortality risk to emphasize the need for expedient treatment (even if symptom acuity has yet to manifest/rise). (Singer et al., 2016)

A bedside SOFA inventory for patients already presenting with more acute symptoms consistent with sepsis was also created. This quick SOFA survey, or qSOFA, has three criteria: altered mental status, systolic blood pressure ≥ 100 mmHg, and respiratory rate ≥ 22 per minute. Using qSOFA, patients meeting any two of these three criteria yielded a predictive validity of 55% in accurately identifying sepsis. For this reason, qSOFA is suggested by Singer and colleagues as an adequate tool for establishing whether subsequent investigation of patient symptoms perhaps consistent with sepsis is necessary. Moreover, qSOFA requires no

lab analyses, making it an expedient, cheap, abridged alternative to an initial, more invasive, and expensive SOFA assessment. (Singer et al., 2016)

Despite the shift to SOFA/qSOFA, a noncontroversial consensus definition for sepsis remains unfulfilled. The 2016 conference conceded that a consolidated, simple definition of sepsis was a lofty goal, given the understanding of etiologic-specific pathophysiology and pathobiology of individual sepsis incidences. The task force charged with pursuing this goal instead offered a prognostic tool for subsequent testing if either a. a patient presented with infection and was already suspected sepsis-positive, or b., a patient exhibited any two of the three qSOFA criteria indicating significant likelihood of mortality absent a health intervention.

1.3 A Critique of the 2016 Sepsis Definition

Though the authors' intentions behind establishing tools for outcome prediction associated with sepsis were good, and though the SOFA/qSOFA scoring systems proved useful tools in a prognostic sense, the authors failed to propose a new and valid sepsis diagnostic tool and definition. In the authors' own words:

> *The agreement between potential clinical criteria (construct validity) and the ability of the criteria to **predict outcomes** typical of sepsis, such as need for intensive care unit (ICU) admission or death (predictive validity, a form of criterion validity), were then tested.*
>
> (Singer et al., 2017)

This statement suggests that the aim of the qSOFA/SOFA tools was for outcome prediction on encountering symptoms consistent with sepsis. The proposed diagnostic framework was less of a valid proposal for establishing a systematic diagnostic decision algorithm, and more a critique

of a past overreliance on inflammation as a valid metric for sepsis. Semantics were modified to eliminate “severe sepsis” as clinically distinguishable from “sepsis”.

The reason that this is a substantive argument against the current Sepsis-3 framework (and for an alternative model) is because clinicians in the intensive care unit (ICU) are tasked with maximizing healthcare outcomes of unstable, acutely sick patients population. This, in contrast with emergency department (ED) clinicians’ responsibilities for health diagnoses and treatment, represents a discrepancy in the intended changes proposed by the Sepsis-3 framework from its predecessors. Moreover, qSOFA and SOFA are recorded as having been validated in an ICU-environment; but sepsis is not limited to the ICU. Prognosis cannot be equated with diagnosis. Effect does not equal cause.

Because medically applied machine learning models and clinical decision support tools are becoming increasingly ubiquitous in the clinical space and given the heavy burden that sepsis represents to the American healthcare system, integration of relevant machine learning models with existing and legacy sepsis diagnostic models deserves serious and immediate attention. Rather than using solely Sepsis-3, clinician gestalt, and electronic health record (EHR) maintained patient health history, the current sepsis diagnostic framework deserves an update. Machine learning can be leveraged to augment the current model of sepsis via comorbidity identification, and ideally, to offer organ-system-specific/context-specific sepsis ‘strain’ diagnosis.

Where symptom-non-disjointness can make sepsis diagnosis convoluted and intangible, machine learning can rectify this issue. Machine learning can augment conventional wisdom via robust calculation of probabilities of disease given the presence or absence of specific features.

Such features are limitless: age, sex, weight, blood pressure, health history, family health history, active comorbidities, living situations, active medications, etcetera. Machine learning thus represents a modern tool capable of delivering the intended outcomes of evidence-based medicine.

This argument is admittedly neither radical nor new. However, the reform of a consensus definition with one that integrates machine learning into its methodology is less common. Existing predictive analyses for sepsis are predicated on conventional wisdom established by criteria outlined in any one of the three modern definitions for sepsis. To the author's knowledge, all machine-learning applications of sepsis prediction suffer from non-generalizable outcomes given limited scope in data sourcing. Populations are often limited to a specific, sometimes predisposed subset of patients whose incidences of disease represent a frequency greater than that of the general public. One study by Calvert and colleagues sought to establish a generalizable machine learning diagnostic tool for sepsis, conceding the same in their methodology (Calvert, Saber, Hoffman, & Das, 2019). An improved understanding that there exist heterogenous manifestations of sepsis has meant that the simple, clear-cut definition as conceived by past consensus conferences may not be attainable. Consequently, machine learning-based approaches are the next logical step in the process.

2. METHODS

The data used for this study were procured from an online repository of clinical trial data and required preprocessing and feature selection prior to predictive modeling.

2.1 DATA SOURCE

The investigation and efficacy of new medical interventions is logged and evaluated by execution of randomized clinical trials, the results of which are added to a repository of clinical trial data in clinicaltrials.gov on completion. Because clinical trials represent new and exploratory analyses on the viability of specific medical interventions there exists an element of risk in their execution. Such risk often manifests in the form of adverse events.

To assess the frequency of target adverse events across multiple clinical trials and to leverage machine learning capability to the data, reformatting into a standardized, numerically indexed scale was required. The LibSVM format is a vectorized representation of data assigning discrete index keys to unique features, and integer, float, or Boolean values representing those features. It is this standardization that potentiates subsequent cross-trial study, given the machine-readable format that it creates. Thanks to this format standardization executed by Tong and colleagues, 30,915 clinical trials were compared with 128,799 unique features between them (Tong et al., 2019). Said features belonged to six separate categories: participant information (discrete values), trial phase (discrete values), serious adverse event (binary/Boolean), other adverse event (binary/Boolean), preexisting condition (binary/Boolean), and interventions (binary/Boolean).

2.2 DATA TRANSFORMATION AND PREPROCESSING

Prior to employment of any machine learning methods, the vectorized data required significant cleaning and preprocessing. Though standardized across multiple clinical trials, the data required further attention for several reasons.

First, all features involving sepsis required removal from the feature space of the data to properly generate target labels denoted by the presence (1) or absence (0) of sepsis outcomes from that clinical trial. A corresponding dataset of feature headers was used to produce a list of sepsis feature headers whose index numbers were referenced and removed while parsing the data.

Second, extraneous features required removal; said features included the clinical trial ID, number of participants, participant median age, total serious adverse events (n=30915), and total, other adverse events (n=30915). The former three features were continuous; the aim of the study was to consider binary/Boolean "presence" or "absence" of features as predictors. The latter two features existed in every clinical trial; the information gained from their inclusion was zero, due to zero variance within each respective feature. None of these features would be useful since the aim of this study was to determine predictive power of factors for sepsis events in a clinical trial context.

Third, prior to model-fitting, the dataset was a textbook case of the so-called "curse of dimensionality" due to its large feature space volume and large ratio of features to rows. Such quantity of features almost certainly makes for model overfitting, rendering outcomes devoid of any clinical significance or meaning. A few methods were investigated and used to lower

chances of over- and under-fitting, namely methods belonging to the families of feature selection, and feature dimensionality reduction alongside regularization.

Below are the raw data as they were received following normalization into this sparse vector format:

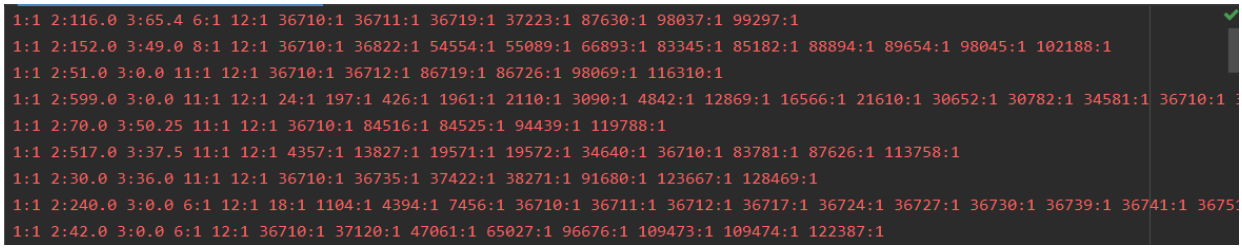


Figure 2. Raw Sparse Data

A separate table mapping each clinical trial feature to a key/index was referenced to search the raw data for features related to sepsis, but excluding “aseptic”:

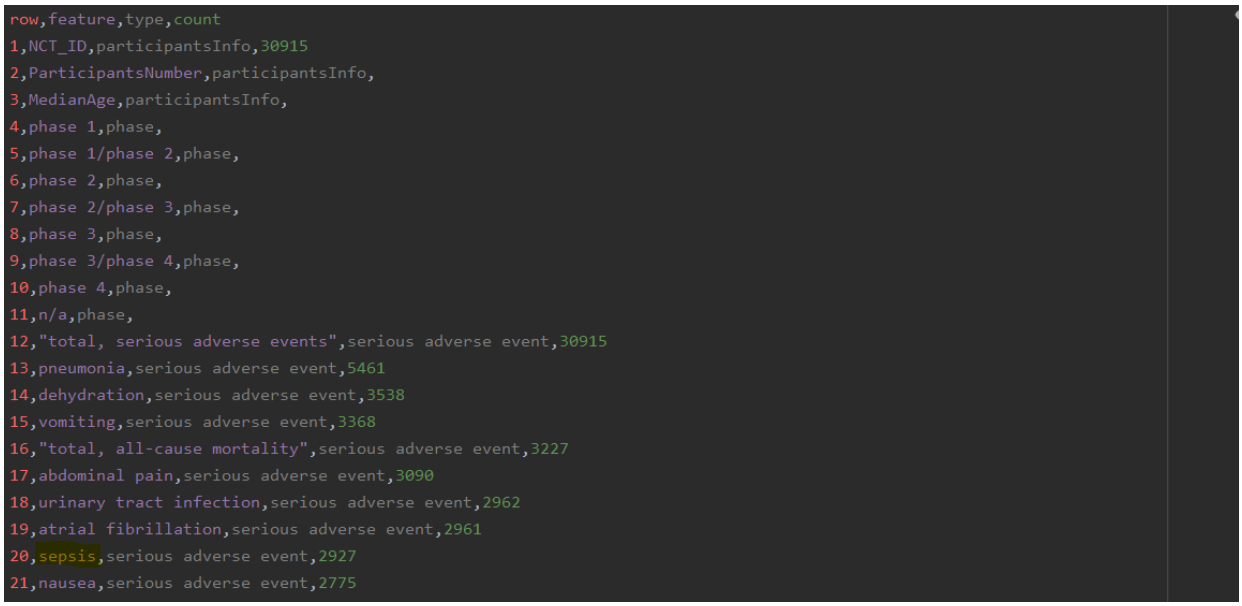


Figure 3. Feature Headers

The following figure is the Python code used to reformat the raw data by excluding sepsis events from the feature space and adding a binary class field indicating presence or absence of sepsis:

```
import pandas as pd

ctDF = pd.read_csv("c:/clinicalTrialSepsisThesis/Feature_Dimensions_Unfiltered.csv",header=None)
hDF = pd.read_csv('c:/clinicalTrialSepsisThesis/Feature_Headers_Unfiltered.csv')

"""
Step 1. Parse thru headers to identify sepsis-related features, excluding 'aseptic'
"""

sDF = hDF[hDF['feature'].str.contains("sepsis") | hDF['feature'].str.contains("septic")]
# n_rows = 427

sDF = sDF['row'][~sDF['feature'].str.contains("aseptic")]
# n_rows = 409 checked hDF for "aseptic" features

# convert pd.df to python list of integers
sepsisIntlist= sDF.values.flatten().tolist()

# list comprehension converting each individual int (index) in the list of ints to a string
sepsisStrlist = [str(x) for x in sepsisIntlist]

string = ":1"

# list comprehension concatenating the ":1" to each sepsis index string
new_sepsisStrlist = [x + string for x in sepsisStrlist]
# new_sepsisStrlist has format ['20:1','68:1',...]
```

```
"""
Step 2. Convert List to String ***
"""
def listToString(list):
    # prior to 'parseline(line)' being called, the data are rows of lists and need converting to a string

    # instantiate an empty string
    string1 = " "
    # concatenate list elements with the delimiter, return the concatenated string
    return string1.join(list)

"""
Step 3. Parse through file by line ***
"""
def parseLine(line):
    # 1 x line-enclosed string holding all indices (has format = ['1:1 2:10.0 3...'])
    line = line.values.flatten().tolist()

    # string, no quotes, no brackets (has format = 1:1 2:10.0 3...)
    stringLine1 = listToString(line)

    # count("12:1") == len(ctDF): would be perfectly correlated w/ sepsis
    stringLine1 = stringLine1.replace(" 12:1","")

    # count("36710:1") == len(ctDF)
    stringLine1 = stringLine1.replace(" 36710:1","")

    # transforms single string to array of strings
    splitLine1 = stringLine1.replace("0    ","").split(" ")


    splitLine2 = splitLine1[3:]

    # new_sepsisStrlist = ['20:1','68:1','...']
    # splitLine2 = ['3:45.9', '10:1','...']

    newLine = ''
    sepsis = any(i in splitLine2 for i in new_sepsisStrlist)
    # if any values in new_sepsisStrlist appear in splitLine2, set equal to sepsis condition

    if sepsis == True:
        # print feature in stringline as long as it's not in new_sepsisStrList
        # place these features in new list object
        newLine = [i for i in splitLine2 if i not in new_sepsisStrlist]

        # concatenate strings in the list
        newLine = ['1', listToString(newLine)]

        # concatenate list of strings to string
        newLine = listToString(newLine)

        # print concatenated string where sepsis condition == True
        print(newLine)

    else:
        # concatenate strings in list
        newLine = ['0', listToString(splitLine2)]

        # concatenate list of strings to string
        newLine = listToString(newLine)
```

```
        # print concatenated string where sepsis condition == False
        print(newLine)

    return newLine

for index, row in ctDF.iterrows():
    newLine = parseLine(row)
```

Figure 4. Preparing the Data

```
0 6:1 36711:1 36719:1 37223:1 87630:1 98037:1 99297:1
0 8:1 36822:1 54554:1 55089:1 66893:1 83345:1 85182:1 88894:1 89654:1 98045:1 102188:1
0 11:1 36712:1 86719:1 86726:1 98069:1 116310:1
0 11:1 24:1 197:1 426:1 1961:1 2110:1 3090:1 4842:1 12869:1 16566:1 21610:1 30652:1 30782:1 34581:1 37172:1 38130:1 3858
0 11:1 84516:1 84525:1 94439:1 119788:1
0 11:1 4357:1 13827:1 19571:1 19572:1 34640:1 83781:1 87626:1 113758:1
0 11:1 36735:1 37422:1 38271:1 91680:1 123667:1 128469:1
0 6:1 18:1 1104:1 4394:1 7456:1 36711:1 36712:1 36717:1 36724:1 36727:1 36730:1 36739:1 36741:1 36751:1 36762:1 36765:1
0 6:1 37120:1 47061:1 65027:1 96676:1 109473:1 109474:1 122387:1
0 5:1 16:1 41:1 54:1 296:1 793:1 15806:1 36711:1 36712:1 36713:1 36714:1 36716:1 36717:1 36720:1 36721:1 36722:1 36723:1
0 8:1 249:1 268:1 619:1 36711:1 36712:1 36714:1 36715:1 36716:1 36718:1 36719:1 36720:1 36721:1 36725:1 36734:1 36736:1
0 11:1 86793:1 86815:1 107589:1 108075:1
0 8:1 13:1 19:1 23:1 25:1 49:1 66:1 77:1 80:1 86:1 106:1 114:1 116:1 130:1 142:1 145:1 186:1 204:1 208:1 274:1 284:1 328
0 8:1 73:1 168:1 36711:1 36719:1 36724:1 36727:1 36751:1 36763:1 86716:1 99127:1 101867:1 108314:1 110678:1
0 6:1 33:1 62:1 363:1 36712:1 36713:1 36714:1 36715:1 36718:1 36721:1 36722:1 36727:1 36734:1 36738:1 36739:1 36740:1 367
0 11:1 13:1 16:1 17:1 27:1 67:1 123:1 156:1 292:1 557:1 735:1 788:1 892:1 2978:1 3190:1 4982:1 13502:1 14334:1 16803:1 17
0 8:1 293:1 36720:1 36731:1 36733:1 36867:1 37077:1 37627:1 37904:1 38268:1 40288:1 50200:1 87100:1 98601:1 108721:1
0 5:1 16:1 36711:1 36712:1 36714:1 36715:1 36717:1 36720:1 36721:1 36722:1 36723:1 36731:1 36733:1 36734:1 36748:1 36755
0 5:1 36711:1 36712:1 36713:1 36714:1 36716:1 36718:1 36720:1 36722:1 36730:1 36743:1 36752:1 36755:1 36758:1 36766:1 367
1 8:1 17:1 27:1 28:1 32:1 47:1 65:1 72:1 103:1 123:1 340:1 399:1 481:1 36711:1 36712:1 36714:1 36715:1 36718:1 36719:1 36
1 5:1 13:1 14:1 15:1 17:1 18:1 21:1 22:1 25:1 27:1 28:1 29:1 30:1 33:1 34:1 36:1 37:1 38:1 39:1 41:1 42:1 44:1 45:1 48:1
0 4:1 36725:1 36791:1 86715:1 98190:1
0 11:1 4045:1 87178:1 95021:1 98037:1 99248:1
0 8:1 22:1 23:1 24:1 30:1 37:1 61:1 70:1 113:1 150:1 161:1 208:1 227:1 255:1 303:1 310:1 581:1 632:1 819:1 825:1 981:1 16
0 8:1 23:1 35:1 157:1 36711:1 36717:1 36718:1 36724:1 36725:1 36734:1 36740:1 36747:1 36773:1 36786:1 36868:1 36889:1 369
0 4:1 36711:1 36715:1 36718:1 36720:1 36722:1 36723:1 36729:1 36730:1 36731:1 36739:1 36744:1 36759:1 36763:1 36765:1 367
0 11:1 94982:1 118503:1 119904:1
0 5:1 36720:1 73787:1 90357:1 90369:1 92807:1 93910:1 120030:1
0 11:1 54018:1 67289:1 79709:1 96257:1 122332:1
0 6:1 198:1 340:1 421:1 775:1 1711:1 36743:1 37148:1 37226:1 86817:1 98036:1 98037:1 98536:1
```

Figure 5. Pre-Feature-Selected Data

2.3. FEATURE SELECTION

Feature selection was important in the case of this study because of the high dimensionality of the feature space (n=128,799) with respect to the number of clinical trials (n=30,915). To that end, filter feature selection methods were used. Wrapper methods were considered, but high feature dimensionality can often render wrapper methods subject to overfitting (Ciortan, 2019). In addition, filter methods were selected as an attempt at pre-modelling data standardization; performance could be based solely on the model, and not on any embedded feature selection parameters within each model. Feature selection helps mitigate issues of overfitting and underfitting the data. Absent feature selection, models may learn the variability in the data too well and consequently overoptimize parameters to fit with the training data only, thus overfitting the model. Conversely, models may be vulnerable to underfitting the data (as is the case with features with low variance). This give-and-take between balancing underfitting (low variance, high bias) and overfitting (high variance, low bias) is the root conflict of the bias-variance-tradeoff.

It should be added that prior to feature selection data are often normalized when the numeric scales between each of the feature columns differ. Data normalization is useful when dealing with continuous features, each with their own distinct numeric ranges. Because all features in the selected feature space were binary (either 0, indicative of “absent”, or 1, indicative of “present”), no such normalization was necessary (Jaitley, 2019).

2.3.1 Variance-Based Feature Selection:

The removal of features with low variance is a necessary step in cleaning the data for successful modeling results. Intuitively, if there exists low variance in features, there exists high similarity among the instances of said features. If a single feature has the same value across many clinical trial instances, the model will learn from these features, to the extent that it underfits the model. The scikit-learn `feature_selection` module and its `VarianceThreshold` method were used to remove all features with zero variance. This method uses a default of 0; thus all features that appeared the same across all 30915 clinical trials were removed. Subsequent retroactive selection of variance thresholds was performed following model execution, and a minimum threshold of variance = 0.01, or 1% (preserving 99% of all variance of the dataset), was ultimately selected. This narrowed the feature space from 128,799 to 754.

2.3.2 Correlation-Based Feature Selection:

Correlation-based feature selection operates on the assumption that "a good feature subset is one that contains features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other" (Hall, 1999). The challenge with a dataset containing semantically different but conceptually similar features (for example, "lesion", "cancer", "tumor", "malignant lesion", etcetera) is to not limit the number of features represented in the final model at the expense of feature uniqueness. For this reason, correlation-based feature selection was used to further select features in the top 20% of f-statistics derived from the ANOVA F-test, implemented by the scikit-learn `f_classif` method

from the 'feature_selection' module. This method returns features' respective F-Statistics and p-values, or the probability that the null hypothesis is true. In other words, p-values are used to determine if mean feature values across positive and negative classes are equal, or, that a feature's presence is independent of the target class. On running the `f_classif` method on the 754 variance-selected features, 306 had p-values of zero. It should be mentioned, however, that all but 13 features had p-values greater than 1%; thus exclusive reliance on p-values for feature importance measurement would have been fallacious. On identifying an f statistic threshold of 20% the correlation-based method selected the final 151 features to be used in model training. These final 151 features were most correlated with sepsis, in that they had the highest ANOVA F-test statistics among the features. Below are the top 10 features as ranked by their f-statistics:

Table 3. Features and F-Statistics

Name	**F-Statistic**
Phase 2/3	11828.65
Pyrexia	9821.41
Pneumonia	9665.61
Myocardial Infarction	8926.04
Cellulitis	8876.59
Atrial Fibrillation	8605.05
Chest Pain	8215.73
Dehydration	7857.79
Back Pain	7687.53
Anemia	7634.04

2.4 DIMENSIONALITY REDUCTION

It should be mentioned that feature selection and dimensionality reduction, though both filtering down the number of features, are two distinct operations. Whereas feature selection excludes selected features from models without changing those features, dimensionality reduction transforms features into a lower dimensional space, by which feature selection automatically follows.

2.4.1 Truncated SVD:

Truncated Singular Value Decomposition (SVD) were initially pursued as an attempt at dimensionality reduction. However, given that the objective of this study was to examine and evaluate discrete clinical trial features as individual predictors of sepsis, and given that truncated SVD consolidates related feature vectors into a summed eigenvector, an implicit loss in granularity of features would have occurred. Moreover, since truncated SVD outputs a predetermined number of vectors, all of which represent generalized eigenvalues of clinical categories, the clinical usefulness of such an output would be difficult to argue in the context of this study. Clinicians are already aware of broad domain-specific covariates of sepsis. The aim here, was to attempt to identify more specific outputs/features.

The output of truncated SVD is graphically represented in a generic, oversimplified form for four model principal components. To note, superscripts indicate correlation with the base feature (such that A^E and E^A communicate some degree of similarity between features A and E). In this simplified diagram, red vectors represent the sum (a principal component) of two vectors found to have high collinearity.

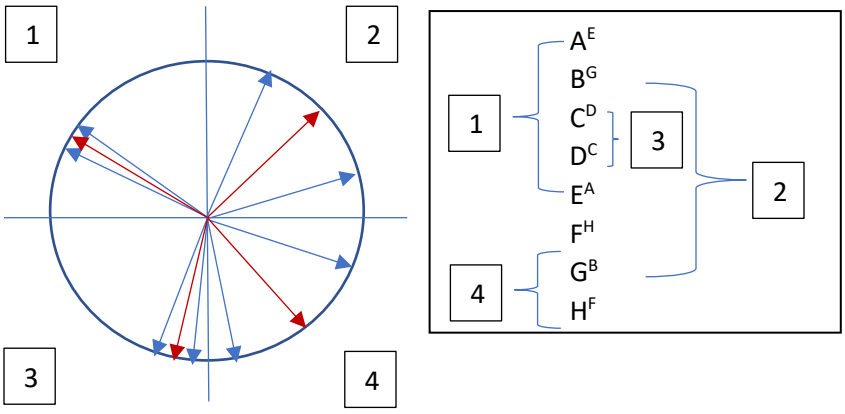


Figure 6. Simplified linear algebra of the Truncated SVD algorithm

Truncated SVD operates on the principle that, in a vector space, a linear map (or a matrix) is a combination of rotated, reflected, scaled, and killed (scaling by 0) vectors. This holds so long as the axes defining that vector space are valid. SVD is a technique leveraging matrix factorization by synthesizing three child matrices from a matrix. (Charan, 2020)

More specifically, if A is a matrix, or linear map, from an n-dimensional vector space V to an m-dimensional vector space W, then A can be considered a product of 3 other matrices, R, D, and S. Here S is an “n x n” rotational matrix with source and target both V; D is an “m x n” diagonal matrix with source V and target W: only non-zero entries are on the diagonal; R is an “m x m” rotational matrix with source W and target W.

First vectors in space V are rotated using S. Second those same vectors are scaled by some constant, and inputted into W by using axes from the map, A. Finally, those output vectors are rotated in W using the R rotational matrix. (Charan, 2020)

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}$$

$V_{dim=n} \longrightarrow W_{dim=m}$

A = R*D*S

$$A = S{:}\begin{bmatrix} s1 & s2 \\ s3 & s4 \end{bmatrix} * D{:}\begin{bmatrix} d1 & d2 \\ d3 & d4 \\ d5 & d6 \end{bmatrix} * R{:}\begin{bmatrix} r1 & r2 & r3 \\ r4 & r5 & r6 \\ r7 & r8 & r9 \end{bmatrix}$$

Figure 7. Simplified Truncated SVD Component Vector Addition

(Charan, 2020)

The SVD technique is similar to another dimensionality reduction technique called Principal Component Analysis (PCA); however, the former operates on raw data matrices, while the latter operates on a covariance matrix (Avila & Hauck, 2017). The incompatibility of PCA with sparse data stems from the fact that it requires operation on an entire matrix (via calculation of a covariance matrix), whereas SVD does not.

Truncated SVD could be useful in subsequent studies to determine organ-system-specific predictors for different pathobiologies of sepsis. However, for the purposes of this study, it is reiterated that truncated SD was abandoned in favor of feature granularity.

2.5 CROSS VALIDATION

Cross validation is typically most useful for datasets with limited numbers of observations (Avila & Hauck, 2017). One type of cross validation called *k*-fold cross validation splits training data into a selected number of equally distributed parts, or folds, assigning one of the folds as a holdout, or test set; the remaining k – 1 folds are used to train the model. For example, a *k*-fold cross-validation where *k* = 100 is a 100-fold cross validation. The model will be iteratively trained using each of the 99 training folds and will test the accuracy of the model by feeding it the hold-out set. Outputted from a 100-fold cross validation model evaluation are 100 *k* scores, which are averaged, and represent the mean model performance. (Kelleher, Namee, & D'Arcy, 2015)

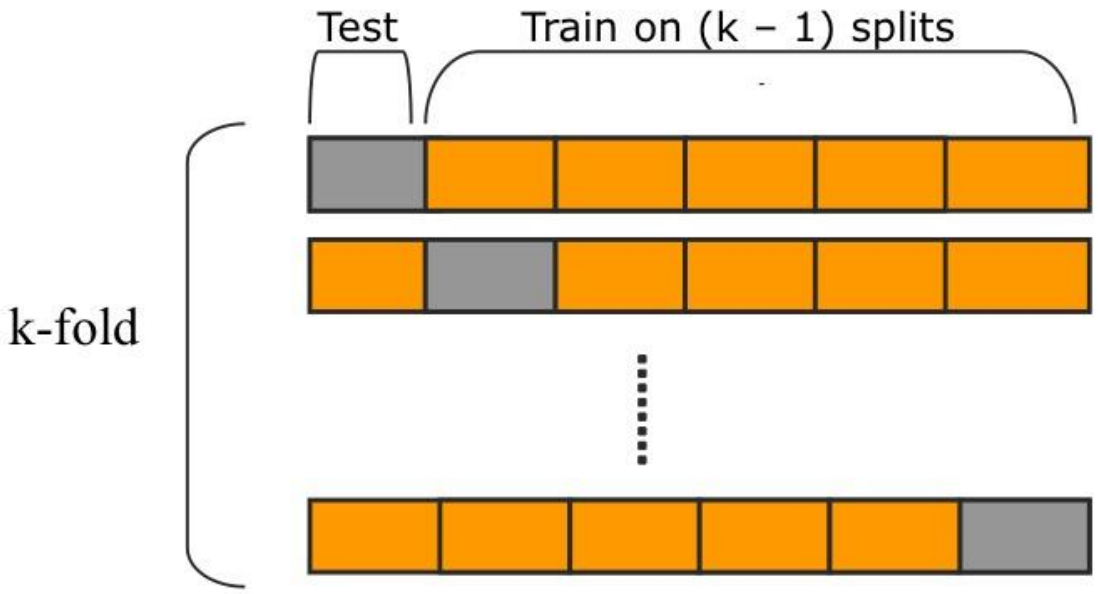


Figure 8. K-fold Cross-Validation

(Kong, 2017)

Given the size of the dataset and given the low computational cost of using the built-in `train_test_split` method of the `model_selection` module, cross validation was

considered but not pursued. This module randomly partitions arrays and/or matrices into training and testing folds of the original dataset. In the case of both the logistic regression baseline model ($n_{features}$ = 128,799) and the feature-selected logistic regression model ($n_{features}$=151), 10-fold cross validation was performed on the dataset. In comparing cross-validated performance with `train_test_split` performance, no statistically significant advantage in the form of better performance was observed in using the former over the latter.

2.6 MODELS

There is a phrase in data science and machine learning that there is "no free lunch when it comes to model selection"; there is no single model that unilaterally performs better across all instances (said instances being data inputs and desired outputs) (Fermin, 2019). Because of this, a number of models were chosen to determine the predictive power of the cleaned clinical trial data. These included logistic regression, support vector machines (SVM) with using linear, polynomial, sigmoidal, and radial basis function (RBF) kernels, naïve Bayes classifier, decision trees, and ensembles of decision trees called random forests. classes.

Following model prediction on subsets of testing data, cross-model performance was evaluated to ultimately select the best-performing model. It should also be added that each of these models belonged to the family of supervised machine-learning classifiers: supervised because the features and classes were labeled; classifiers because the object is to determine predictive power of datasets with feature labels, and subsequently evaluate and classify the predictive power of each feature.

2.6.1 Logistic Regression

Logistic regression (LR) is a binary classification algorithm that assigns a class to a categorical feature via application of a 50% probability (Geron, 2019). LR calculates the weighted sum of the entire input feature space and outputs its biased logistic, scaled between 0 and 1, where an output greater than 0.50 indicates positive association with the target variable, and an outcome less than 0.50 indicates a negative association with the target variable. More specifically, LR applies the sigmoid function [0,1]:

$$\sigma(x) = \frac{1}{(1+e^{-X})}$$

or graphically represented:

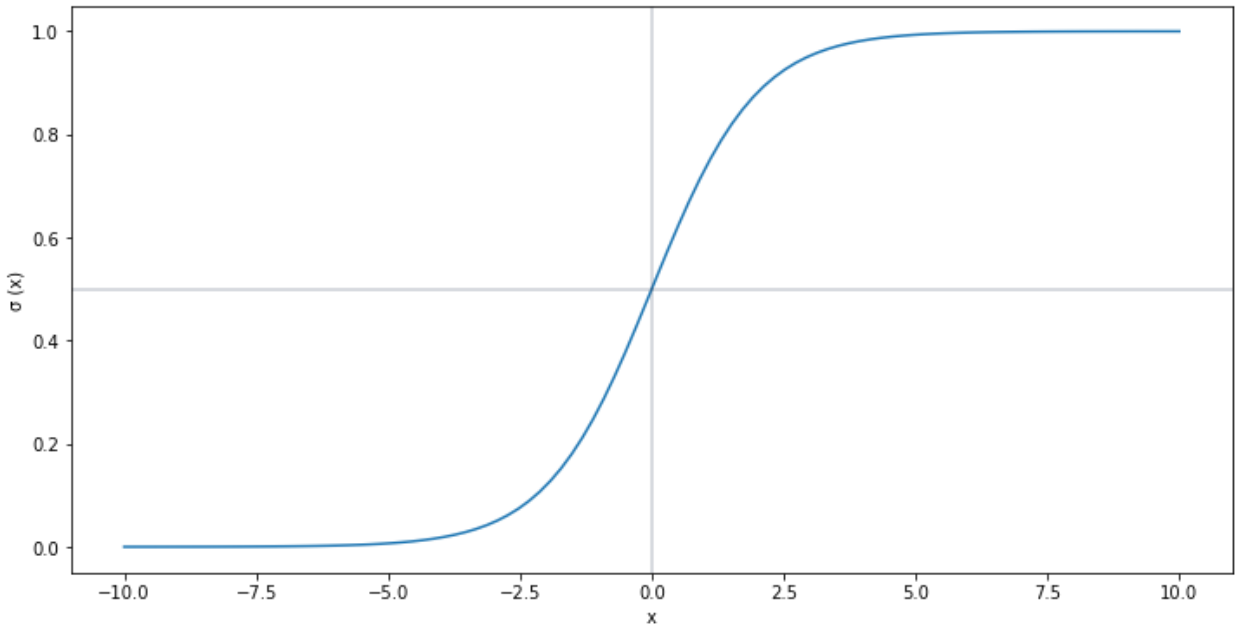


Figure 9. Sigmoid Function

(sklearn.linear_model.LogisticRegression, 2020)

Within this scale [0.0, 1.0], all logistic regression models estimate probabilities that some feature/instance belongs to a class, or that:

$$\hat{p} = h_\theta(\text{x})$$

Where $\hat{p}$ equals the calculated probability that the instance, x, belongs to the category, h_θ. In the context of this case study:

$$\hat{y} = \begin{cases} 0 \; if \; \hat{p} \; < 0.5 \\ 1 \; if \; \hat{p} \; \geq 0.5 \end{cases}$$ that a feature is consistent with sepsis

where $\hat{y}$ is the classification produced by the model. Thus it stands to reason that if an estimated probability that a feature belongs to the sepsis class equals 0.6, there is a 60% chance that a feature will predict sepsis.

Though logistic regression produces such probabilities according to this non-linear sigmoidal function, it is still considered a linear model because on solving $\sigma(x) = \frac{1}{(1+e^{-x})}$ for x, where the sigmoid, σ(x), equals *p* (or the probability that an observation belongs to the response variable/class):

$$p = \frac{1}{(1+e^{-X})} \; (sigmoid \; function)$$

$$e^{-X} = \frac{1}{p}$$

$$e^{-X} = \frac{1}{p} - 1$$

$$e^{-X} = \frac{1-p}{p}$$

$$e^{X} = \frac{p}{1-p} \leftrightarrow \frac{p}{q}$$

$$X = \log \frac{p}{q} \; (logit \; function)$$

(Klosterman, 2019)

the resulting equation for the log odds, or odds ratio, represents a probability given a linear combination of all features in the feature space. Thus, because X, or the aggregate linear combination of features when the response variable equals the logit function, logistic regression is a linear model. In other words, given a feature space of size n:

$$p = \frac{1}{(1+e^{-(\theta_0+\theta_1 X_1+\theta_2 X_2+\cdots\theta_n X_n)})} \quad \text{(sigmoidal logistic regression)}$$

$$\theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots \theta_n X_n = \log \frac{p}{q} \text{ (log odds logistic regression)}$$

Because the sigmoid equation can be *unilaterally* generalized with such a transformation into the logit function, or, into a $y = mx + b$ form, it is proven that logistic regression is linear.

(Klosterman, 2019)

2.6.1.1 REGULARIZATION

Lasso and Ridge regularization methods are two methods that assign penalty, or cost, for having larger coefficients in a fitted model. In short, these methods assign cost, or penalties for predicting values incorrectly, and in doing so are integral in parameter optimization for returning the least "costly" model. By doing this, the model learns from an inputted training set of data for model fitting that can generalize to new data on being asked to predict outcomes. The log-loss function is one such cost function used in scikit-learn for penalty assignment and model fitting in a number of models, but most notably logistic regression:

$$\log loss = H_p\,(q) = \frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log(p_i)) + (1 - y_i) \cdot \log\,(1 - p(y_i))$$

(Klosterman, 2019)

where n is equal to the number of samples, y_i is equal to the actual label of a sample of index, *I,* and p_i is the probability that a sample at index I belongs to the target class (or $p(y_i)$ = 1). By

optimizing model parameters, the response variable log odds and individual features' log odds are calibrated to the other to minimize the cost function. Log loss is also called the cross-entropy function, or simply the "difference between two probability distributions for a given random variable or set of events" (Brownlee, 2019).

Two extrapolations on the log loss function are lasso (L1) and ridge (L2) regularization methods. Both methods leverage the log loss function, but each use a different term to minimize the cost function. The L1 regularization appends the log loss with the 1-norm:

$$\text{lasso penalty log loss} = \sum_{j=1}^{m} |\sigma_j| + \frac{C}{n} \sum_{i=1}^{n} (y_i \cdot \log(p_i)) + (1 - y_i) \cdot \log\big(1 - p(y_i)\big)$$

and the L2 regularization appends the log loss with the 2-norm.

$$\text{ridge penalty log loss} = \sum_{j=1}^{m} \sigma_j^2 + \frac{C}{n} \sum_{i=1}^{n} (y_i \cdot \log(p_i)) + (1 - y_i) \cdot \log\big(1 - p(y_i)\big)$$

(Klosterman, 2019)

Except for the operations performed by the first term, the two regularization models are identical. The key difference is that L1 includes the sum of absolute values of coefficients between m different features, whereas L2 includes the sum of squares. With respect to performance, L1 can be used as a feature selection method if given a coefficient equal to exactly zero, as this assignment eliminates the feature. L2 penalties do not eliminate features given a coefficient value of zero.

In the case of this study, binary logistic regression was used, given that the response variable accounted for two possible outcomes: sepsis, and not sepsis. Within the `LogisticRegression` method, cost penalties L1 or L2 are specified.

2.6.1.2. Solvers

The `LogisticRegression` method also accepts a solver parameter. Solvers find parameter weights for further minimizing the cost function, previously specified. Used in this study were liblinear, sag, and saga solvers.

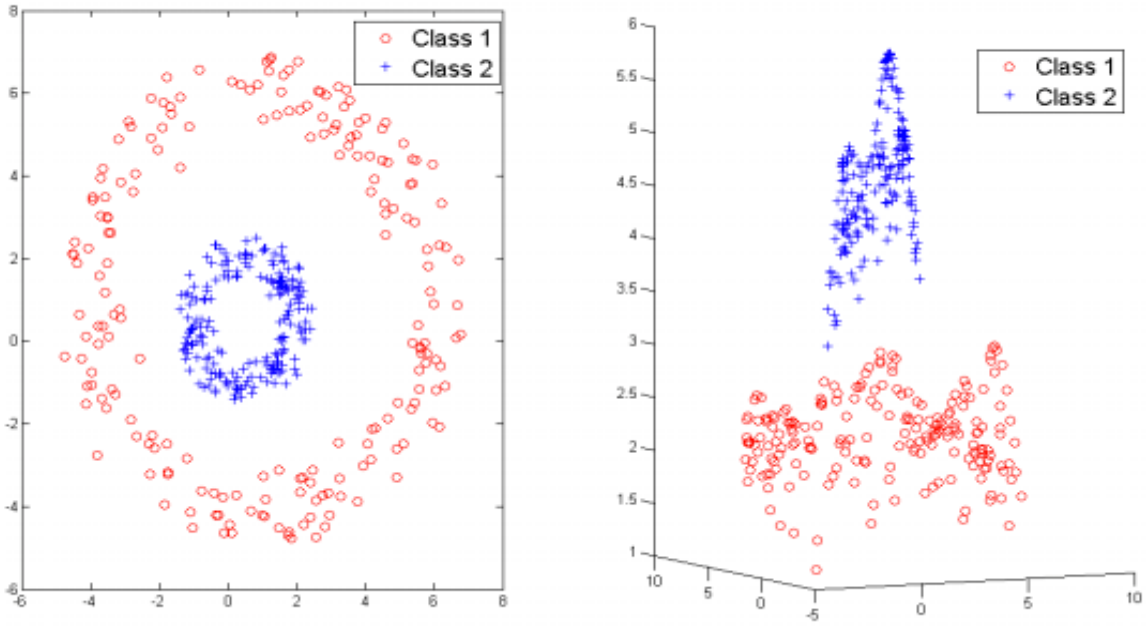
The liblinear solver minimizes the cost function from a single direction at a time; given this detail, logistic regression with this solver performed noticeably slower than the same models using gradient descent. Sag, or the stochastic average gradient descent solver, abbreviates parameter calculation by randomly sampling in-cache gradient values, but is limited to L2 regularization. Saga is a variant of Sag, but instead is compatible with L1 regularization, which allows for input of sparse data. Both the sag and saga solvers are optimized for larger datasets. (sklearn.linear_model.LogisticRegression, 2020)

2.6.2 Support Vector Machine (SVM)

Support vector machines (SVM) are another form of classification algorithm but do not require linearity for classification. SVMs determine and optimize two-, three-, or multi-dimensional hyperplanes as modes of classification, as well as decision boundaries in cases where data points are not unilaterally/linearly separable. Ultimately SVM aims to separate datapoints by some maximum distance from a hyperplane called a margin (Kelleher, Namee, & D'Arcy, 2015). The support vectors are those datapoints.

In cases where separability is indeed not linear, additional functions called kernels are applied for decision function specification. Kernels ultimately act as operators that apply some weight to the data, transforming the distances between datapoints in the aggregate so a

hyperplane is more easily found. Weights may be unilaterally applied (linear) or nonlinear, depending on the kernel function. This is simplified for the sake of exposition below, where a two-dimensional feature space (where data are non-linearly separable via hyperplane) is kernelized, or projected, into a three-dimensional feature space (where data are linearly separable via a hyperplane with maximum margin):



(Fletcher, 2009)

Figure 10. Kernel Trick for Hyperplane Identification

Kernels provided by sklearn include a linear, polynomial, radial basis function (RBF), or sigmoid (similar to logistic regression). Below is the famous Iris dataset, operated on by multiple SVMs-based kernels:

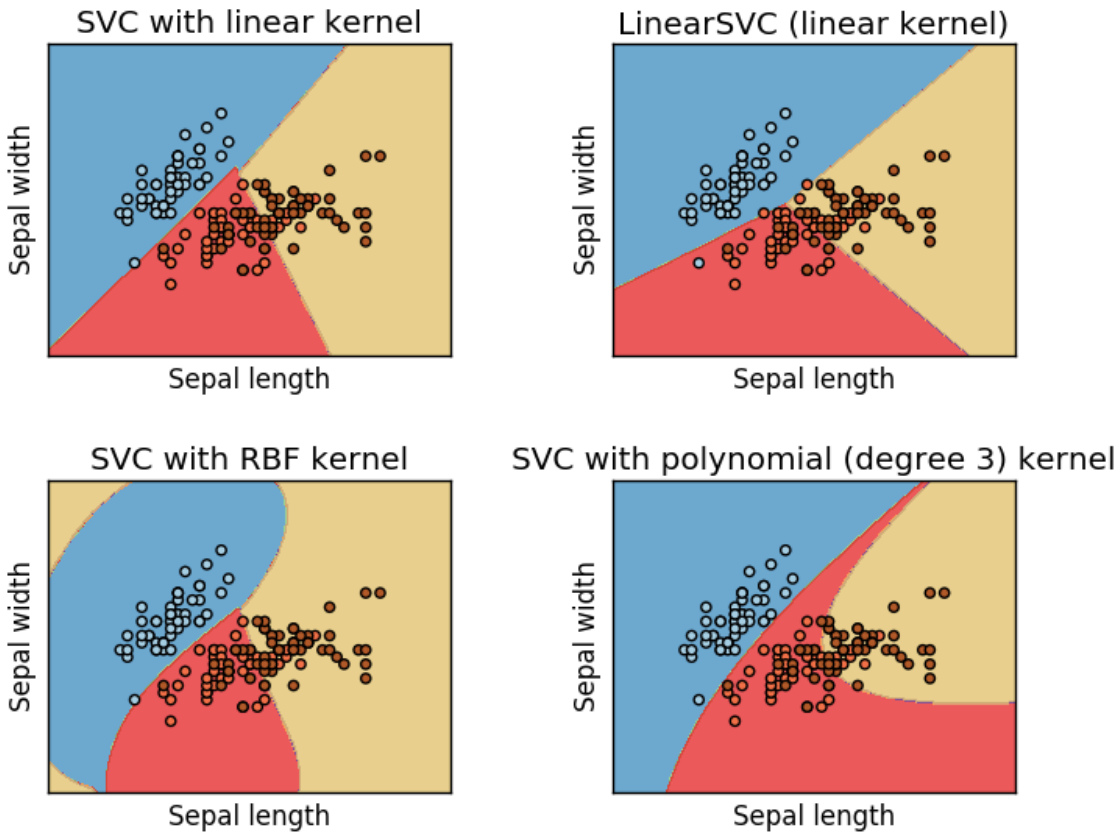


Figure 11. Various SVM Kernel Applications on Iris Dataset

(sklearn.svm.LinearSVC, 2020)

2.6.2.1 Linear Kernel

The linear kernel in a support vector machine is used when data is linearly separable. This kernel is most typically used when the data have many features and two classes, making this kernel ideal for the dimensions of the dataset. A linear hyperplane can be a line in two dimensions, or a plane in three dimensions, and can be represented in a standard $y = mx + b$ format, with dimensions as parameters:

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n = 0$$

2.6.2.2. Radial Basis Function Kernel

The Radial Basis Function (RBF) kernel is a kernel used under similar circumstances to the linear kernel, but instead takes Euclidean distance between data points into account. The RBF Gaussian kernel is as follows:

$$k_G(x, x') = \exp\left(-\frac{d(x, x')^2}{2\sigma^2}\right)$$

where σ equals a parameter, and d is the Euclidean Distance (or direct-line distance) between the two data points x and x' (Vert, Tsuda, & Schölkopf, 2004).

2.6.3 Naïve Bayes

Naïve Bayes classification is another classification algorithm that naively assumes feature independence in class prediction. Below, the simplified Bayes' Theorem:

$$P(A|B) = \frac{\left(P(B|A) * P(A)\right)}{(P|B)}$$

As an example applied to the context of this study, if sepsis is related to blood pressure, then, using Bayes' theorem, a person's blood pressure can be used to more accurately assess the probability that they have sepsis than can be done without knowledge of the person's blood pressure.

A Naïve Bayes assumption states that features are conditionally independent of each other given some response variable. Or:

$$P(X1|X2, Y) = P(X1|Y)$$

or in the case of the 151 features from the dataset:

$P(Sepsis|X_1,X_2,\ldots,X_{151})$

and applied to Bayes' Theorem:

$$P(sepsis|X_1,\ldots,X_{151}) = \frac{P(X_1|sepsis)P(X_2|sepsis)\ldots P(X_{151}|sepsis)P(sepsis)}{P(X_1)P(X_2)\ldots P(X_{151})}$$

The denominator remains unchanged and can thus be eliminated when determining class outcome:

$$P(sepsis|X_1,\ldots,X_{151}) \propto P(X_1|sepsis)P(X_2|sepsis)\ldots P(X_{151}|sepsis)P(sepsis)$$

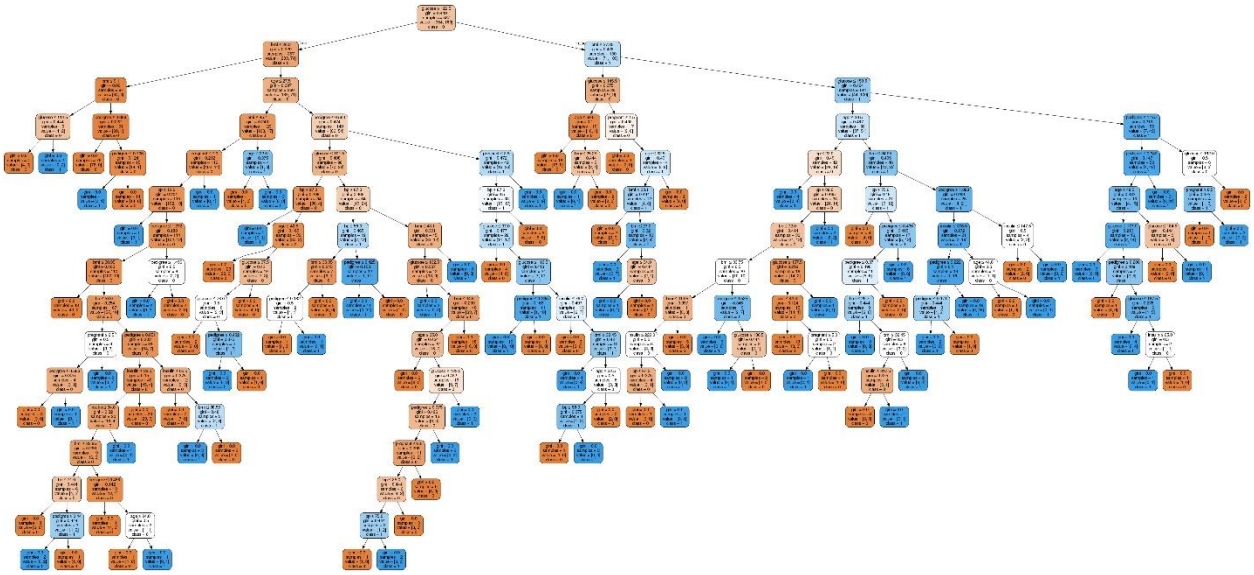
Then, given predictors, Sepsis can be identified:

$$sepsis = argmax_{sepsis}P(sepsis)\prod_{i=1}^{151}P(X_i|sepsis)$$

Naïve Bayes classifiers uses maximum likelihood estimation for parameter estimation, taking into account conditional probability and prior probability. There exist multiple types of Naïve Bayes classifier. The two used in this study included Bernoulli and Gaussian Naïve Bayes. The former uses Boolean features for class prediction (0 and 1), and the latter uses continuous features.

2.6.4 Decision Tree

Decision trees are recursive decision structures that seek to maximize quantifiable information gain by identifying the most informative feature for each node and applying subsequent decision splits to each subsequent feature. These algorithms are instantiated at an aggregate "bin" of all possible features, called a root node, representative of the population being sorted. Each subsequent decision point below the root node is called a node, and final decisions where no further splits are necessary are called terminal nodes.



(Datacamp, 2019)

Figure 12. Decision tree: unpruned

Decision trees are among the most popular machine learning algorithms because they are almost immediately interpretable both in rationale and practice; the issue of the machine learning "black-box" is less likely to apply to decision trees. Additionally, they are capable of both regression and classification tasks (predicated on whether the machine learning problem is concerned with continuous or categorical features), making them versatile and good baseline models against which to compare other non-tree-based models or ensemble tree models.

Decision splits are determined by a number of different decision rules, but in the case of this study were Entropy and Gini rules. The former determines the variety of possibilities, or disorder, of a target variable. The latter is a measure of the impurity, or the rate at which a randomly chosen feature predicts the wrong class.

A few popular decision trees algorithms include the Iterative Dichotomizer 3 (ID3), CART (Classification and Regression Trees), CHAID (Chi-squared Automatic Interaction Detection for classification tasks). A number of derivative machine learning algorithms have come from

decision tree classifiers including fast and frugal trees (minimally deep/maximally shallow decision trees designed to aid decision flows in professional spaces where the bulk of decision-making can be distilled down to a few crucial steps), and extrapolative bootstrapping models like random forests and gradient boosted decision trees.

Decision trees are at a disadvantage when used on their own or when working with data of high dimensionality. If not for the feature selection and engineering methods employed, the high variance of the data would have caused a non-generalizable model. Despite this, a decision tree model was applied to the dataset for the sake of exposition re: the continuum of tree-based classifiers ranging from decision trees to random forest classifiers (and perhaps in subsequent study, gradient boosted tree algorithms) and its clear representation of knowledge.

2.6.5. Random Forests

Random forests bagging ensembles, or collections/forests of decision trees that have been bootstrapped. Bootstrapping is the process of resampling the training dataset in parallel with model fitting and replacing poorly performing samples. Bootstrapping aggregates all these inputs via a number of different techniques, but sklearn compiles a list of all predicted probabilities for each feature, selecting the feature with the highest probability as an output. On generating some specified number of decision trees with the `n_estimators` parameter, trees are averaged, and a prediction is outputted (sklearn, 2020).

Random forests are greater than the sum of their parts, in that their strength lies in the numbers of their constituents. Outputs are committee-based and represent a collective "agreement" among the group.

3. OBJECTIVES

The objectives of the study were three-fold. First, the study was a practice in data manipulation and reformatting, which required in-depth understanding of data structures and data types. Before the data were ready for machine learning, there existed multiple indexed sepsis events that required removal from the feature space, but whose presence or absence required denoting.

Second, this thesis represented an introduction to medically applied machine learning as well as feature and model selection.

Third, this thesis aimed to contribute to two conversations. The first, that the continued/further integration of machine learning applications into the clinical space could supplement modern evidence-based medicine best-practices for disease diagnosis, prediction, and prevention/intervention, thus potentiating maximal health outcomes. The second, that a new sepsis definition should be pursued, ought to incorporate machine learning into its underlying framework, and should require explicit diagnostic criteria including covariates of disease, rather than settling on ED/ICU-gathered prognosis outcomes self-described as markers for diagnosis. It is reiterated that use of prognostics as diagnostics for sepsis is dangerous, especially when the most dire prognoses are accepted as criteria for sepsis to begin with.

4. RESULTS

Table 4. Optimized Model Results Table

Algorithm	Tags	Feature Selected Model				
		Precision	Recall	F1-Score	AUROC	Accuracy
Logit	No Sepsis	0.94	0.97	0.95	0.79	**0.92**
	Sepsis	0.77	0.61	0.68		
SVM	No Sepsis	0.94	0.97	0.95	0.78	**0.92**
	Sepsis	**0.78**	0.58	0.67		
Naïve Bayes	No Sepsis	0.95	0.91	0.93	**0.83**	0.86
	Sepsis	0.56	0.71	0.62		
Decision Tree	No Sepsis	0.92	0.97	0.94	0.73	0.91
	Sepsis	0.71	0.51	0.59		
Random Forest	No Sepsis	0.93	0.97	0.95	0.78	**0.92**
	Sepsis	0.76	**0.72**	**0.74**		

TP = True Positive
TN = True Negatives
FP = False Positive
FN = False Negative

Precision = $\frac{TP}{TP+FP}$ *or* $\frac{TP}{Identified\ Positives}$ Recall = $\frac{TP}{TP+FN}$ *or* $\frac{TP}{Actual\ Positives}$

F1 = $2 * \frac{Precision*Recall}{Precision+Recall}$

The results table includes precision, recall and f1 scores for each of the feature-selected selected models. The random forest model performed best and identified phase 2/phase 3, phase 3/phase 4, sleep apnea syndrome, deep vein thrombosis, dyspnea, atrial Fibrillation, chronic obstructive pulmonary disease, cancer pain, acute cholecystitis, peritonsillar abscess as the top ten most predictive features.

The best performing models are summarized with the following AURCOCs, and normalized/non-normalized confusion matrices.

4.1. Logistic Regression: Sag Solver

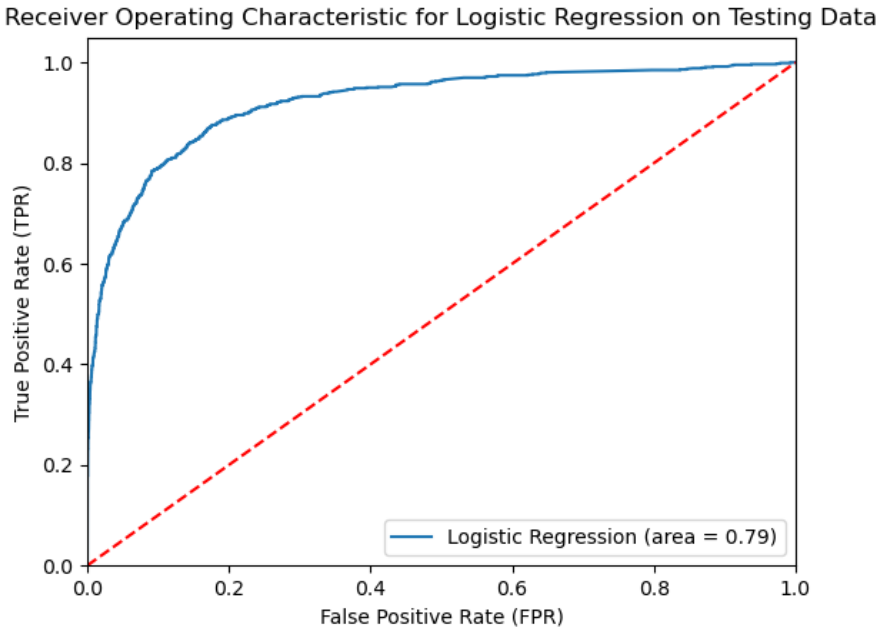


Figure 13. AUCROC For Logistic Regression Sag Solver

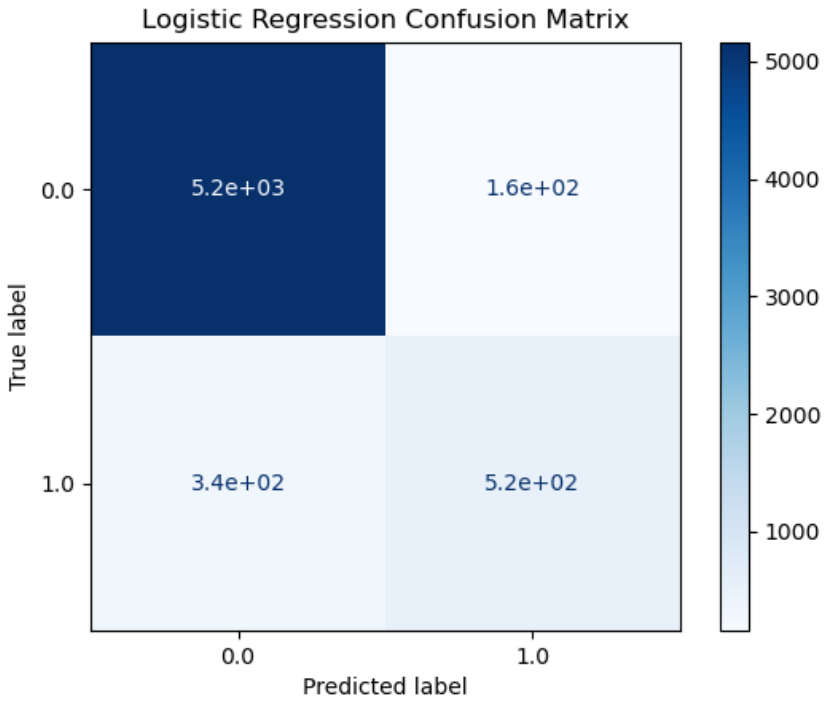


Figure 14. Logistic Regression Sag Solver Confusion Matrix

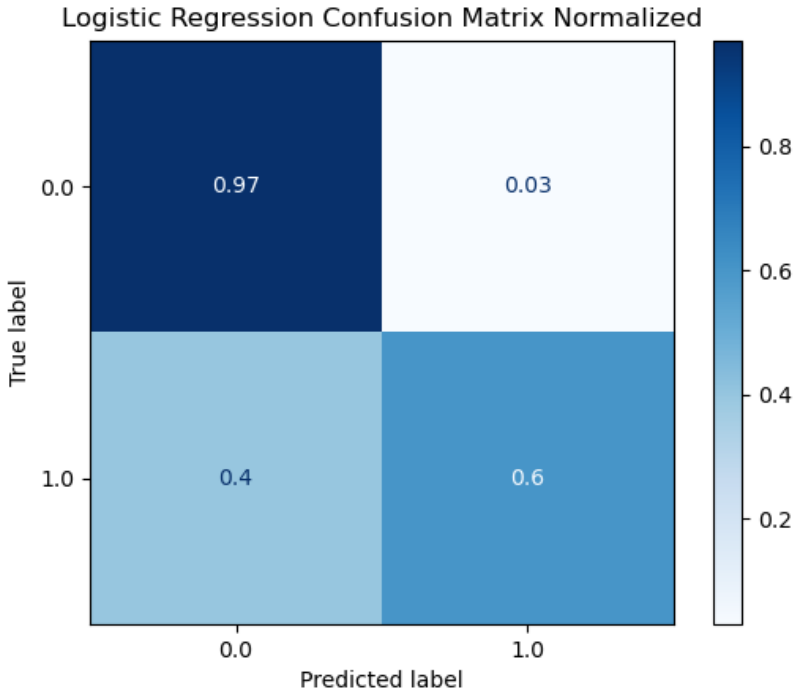


Figure 15. Normalized Confusion Matrix for Sag Solver

4.2. Support Vector Machines: Linear Kernel

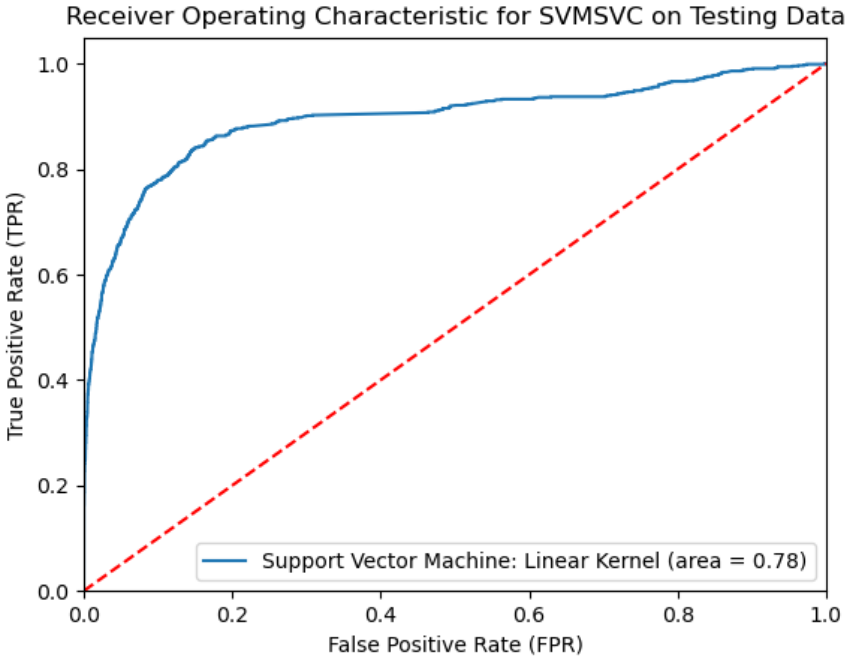


Figure 16. AUCROC for SVM.SVC Linear Kernel

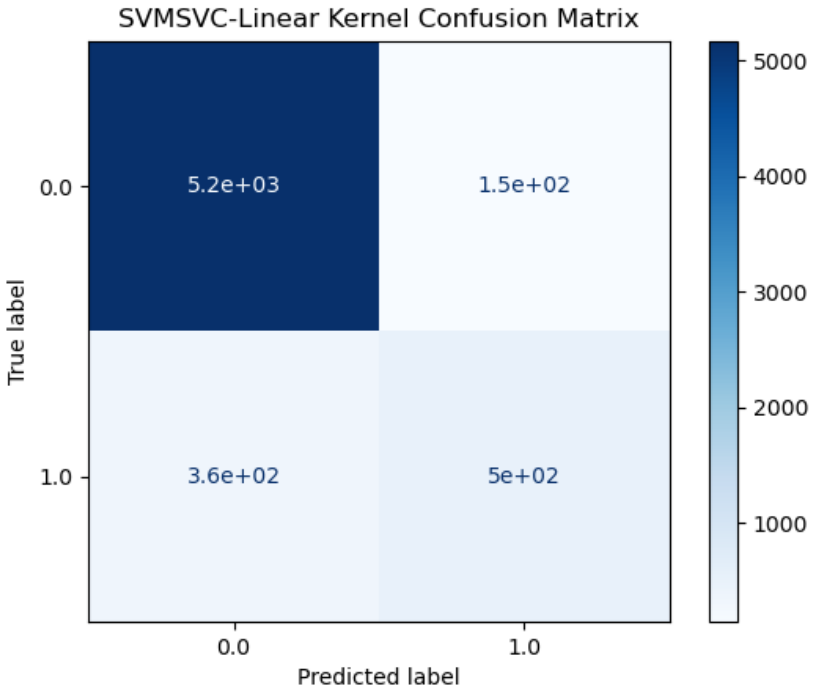


Figure 17. SVM.SVC Linear Kernel Confusion Matrix

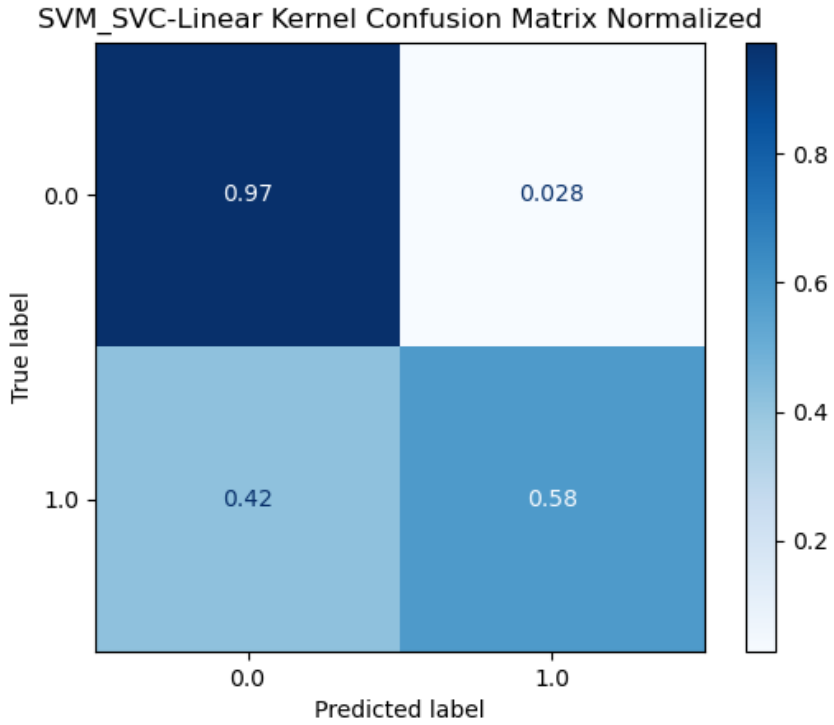


Figure 18. Normalized SVM.SVC Linear Kernel Confusion Matrix

4.3. Naïve Bayes Classifier: Gaussian Classifier

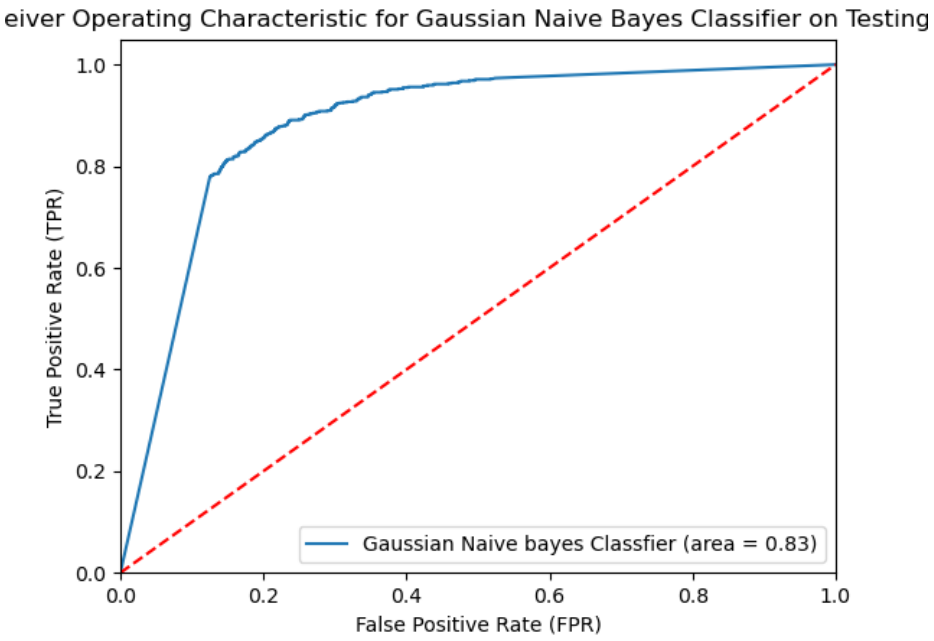


Figure 19. AUCROC For Naïve Bayes Gaussian Classifier

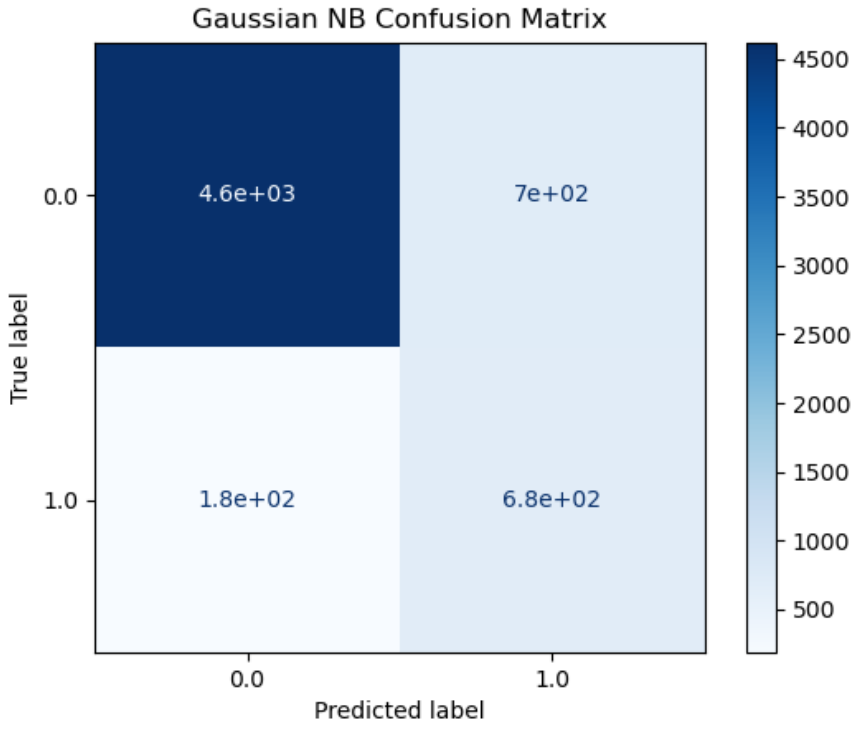


Figure 20. Naïve Bayes Gaussian Classifier Confusion Matrix

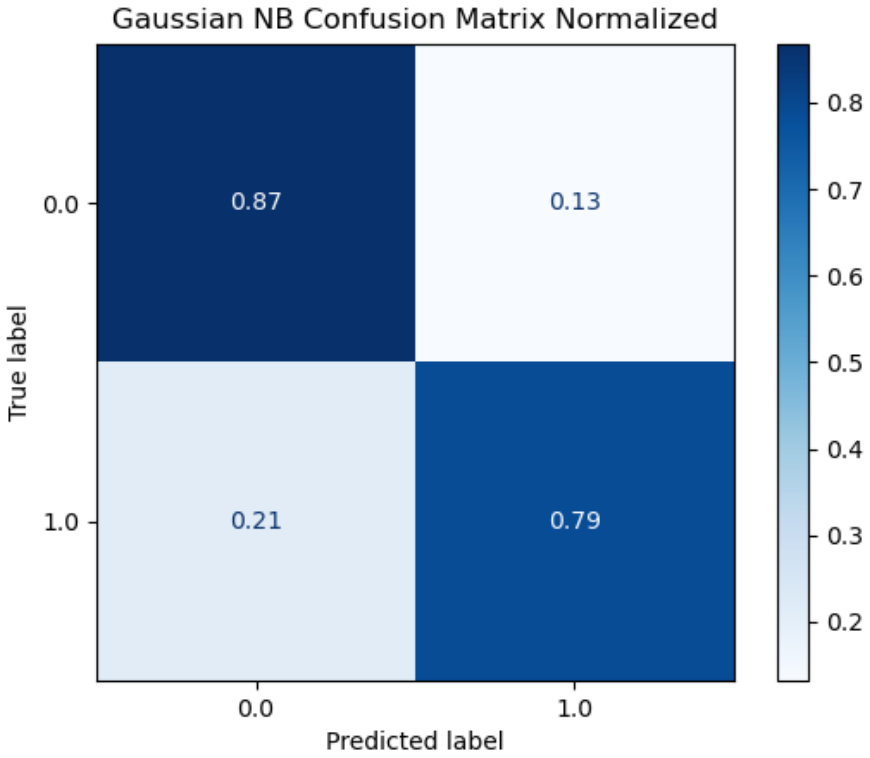


Figure 21. Normalized Naïve Bayes Gaussian Classifier Confusion Matrix

4.4 Decision Tree: Gini Depth 10

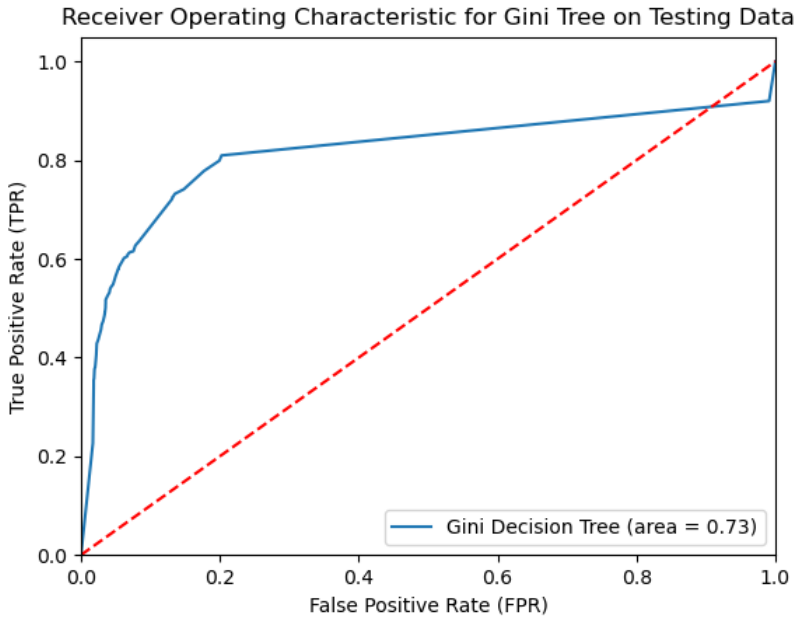


Figure 22. AURCOC for Gini Impurity Depth 10 Decision Tree

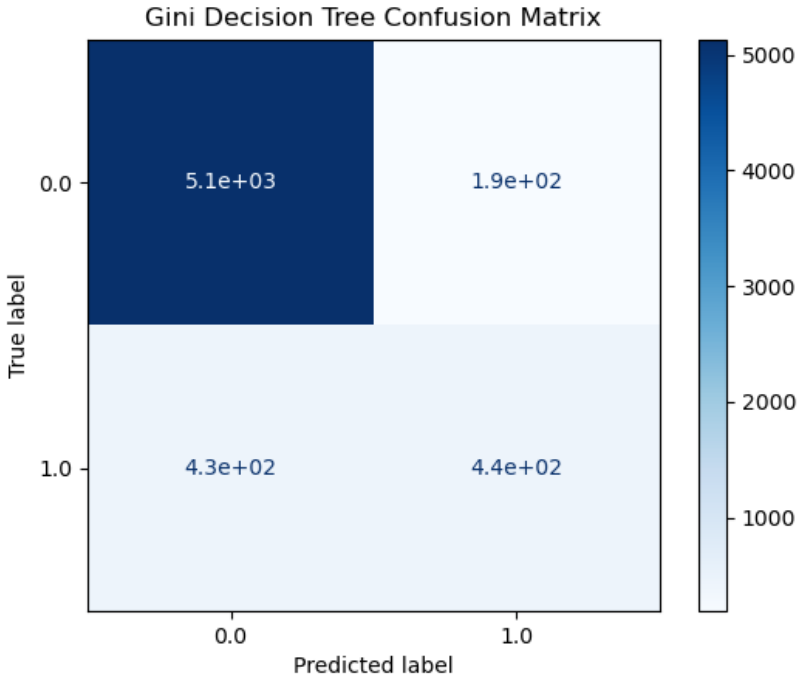


Figure 23. Gini Impurity Depth 10 Decision Tree Confusion Matrix

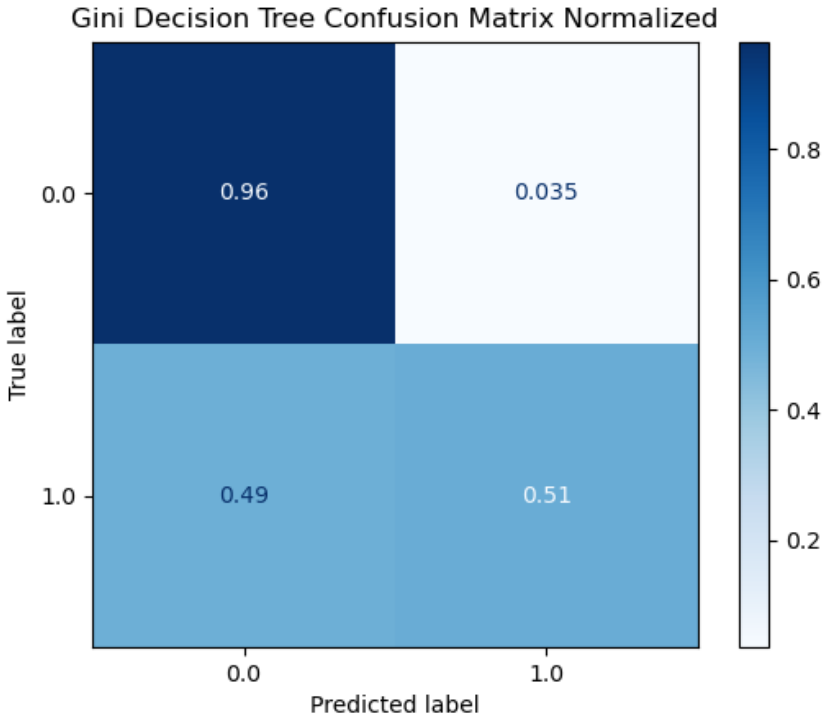


Figure 24. Normalized Gini Impurity Depth 10 Decision Tree Confusion Matrix

4.5. Random Forest: Gini Impurity

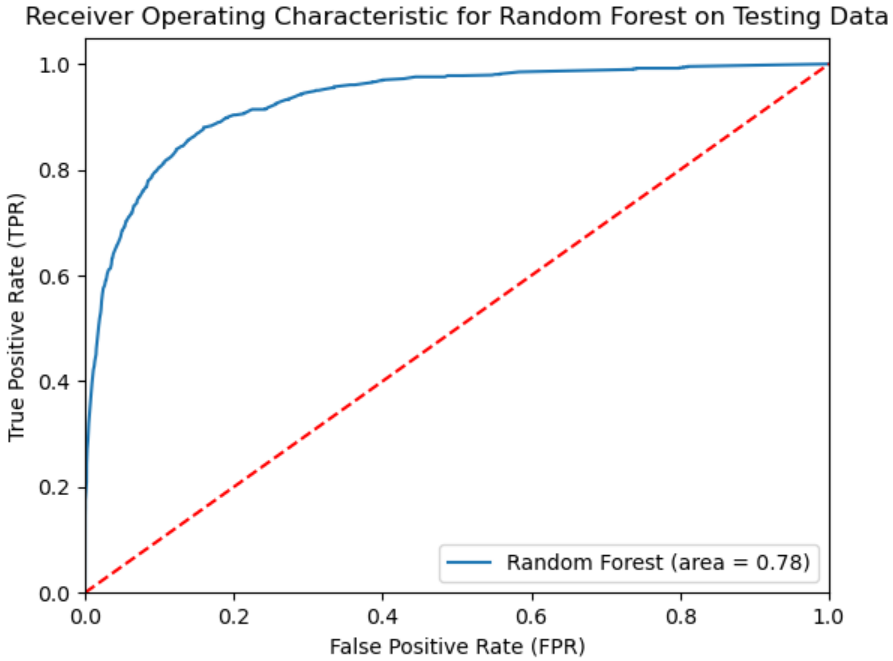


Figure 25. AUROC for Random Forest Gini Impurity

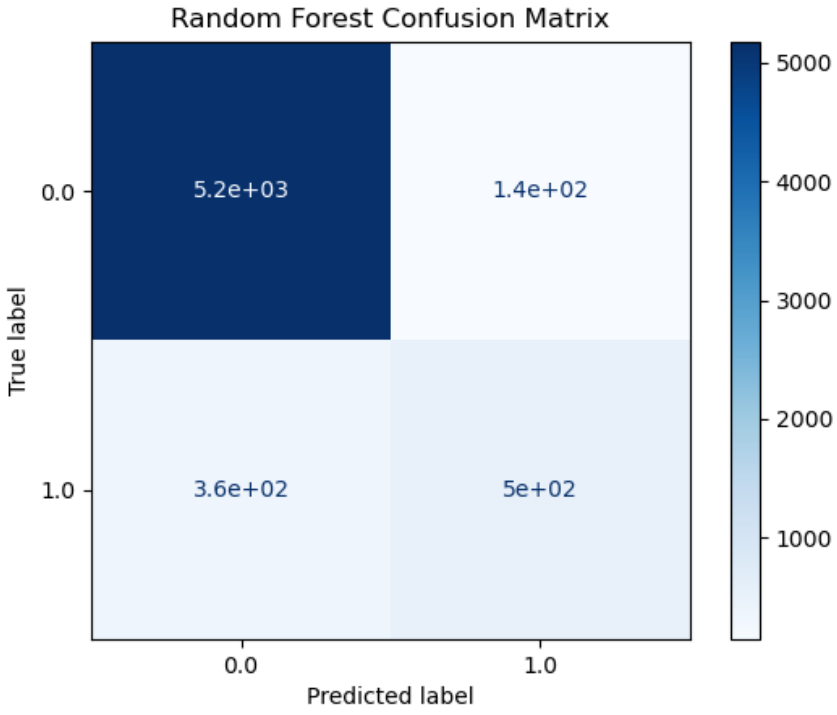


Figure 26. Random Forest Gini Impurity Confusion Matrix

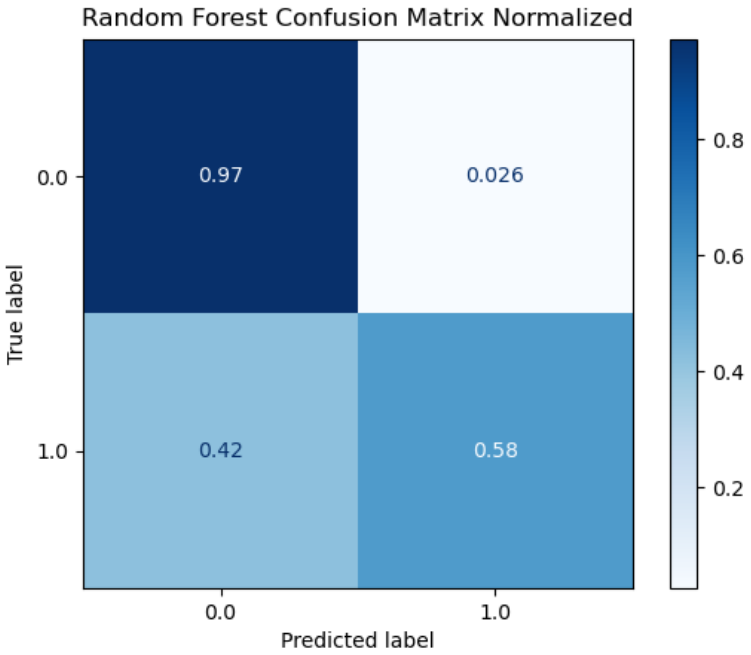


Figure 27. Normalized Random Forest Gini Impurity Confusion Matrix

5. DISCUSSION

Given that only 4,367 of the 30,915 (14%) of clinical trials reported any incidences of sepsis, the issue of class imbalance was significant. In machine learning, the most accurate models are those trained on data with classes that are near equal in distribution. In the case of binary response variables (0 for no sepsis, 1 for sepsis), this optimal distribution would have had 15,458 incidences each of sepsis and no sepsis.

Accuracy scores were likely bloated due to class imbalance as well. Moreover, an over-reliance on accuracy as a robust assessment of the models' performances in the instance of this dataset would be misleading. Because models predict both sepsis-related (1) and non-sepsis-related (0) events, and because the ratio of non-sepsis:sepsis in the data is roughly 7:1, accuracy figures are too heavily influenced by the models' specificities (true negative rate). The same weight is being given to specificity as sensitivity, despite the class imbalance.

A subsequent study could work to further improve preprocessing steps to target a 1:1 ratio class events using a technique called the Synthetic Minority Over-sampling Technique, or SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002). This method synthetically generates new samples consistent with the minority class distribution and is available in the `imbalanced_learn` library's `SMOTE` class. Further investigation into boosted ensemble classifiers like gradient boosted trees/models, Adaboost, or XGBoost could be a means of applying the subject matter of this study to more relevant and modern models.

Finally, as the title suggests, the approach to sepsis classification is naïve, in that sepsis events are lumped into a single feature. Given that that modern understanding of the pathobiology and pathophysiology of sepsis suggests its mechanism may be organ-system

specific, the follow-up study could assign nominal or ordinal categories to different sepsis categories.

6. CONCLUSION

The healthcare industry is experiencing a paradigm shift in how it identifies, describes, predicts, intervenes, and prevents major adverse health outcomes thanks in large part to machine learning models. Constant revision of modern machine learning techniques guarantees improved results. When applied to healthcare, such improved results are assumedly countless.

It is difficult to conceive of a future where machine learning will not continue to impact healthcare, if not already doing so. Machine learning is becoming increasingly common in the healthcare data space including but not limited to: clinical decision support tools; health vitals analytics and prediction made possible by data generation and aggregation from wearable health devices, electronic health record data, and insurance health claims data; image processing of medical images for early detection of disease; pharmaceutical development and design via machine learning aided discovery of prognostic biomarkers (Vamathevan et al., 2019); and this list continues to grow.

In this study a suite of models was used to identify predictors of sepsis from 128,799 features distributed among 30,915 unique clinical trials. Thanks to the open-source nature of machine learning libraries like scikit-learn, and a readiness by the data science and machine learning community for knowledge transfer, this project has been an illuminating practice in machine learning applications and platforms for discourse. Though there is no turnkey solution for health outcome maximization, artificial intelligence and machine learning represent a tinkerer's paradise for iterative learning and hypothesis testing that can only make a future with fewer chronic illnesses more likely.

REFERENCES

1.10. Decision Trees. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/tree.html

1.11. Ensemble methods. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/ensemble.html

1.4. Support Vector Machines. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/svm.html

1.9. Naive Bayes. (2020.). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/naive_bayes.html

3.2.4.3.1. sklearn.ensemble.RandomForestClassifier. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Avila, J., & Hauck, T. (2017). *Scikit-learn cookbook: over 80 recipes for machine learning in Python with scikit-learn* (2nd ed.). Packt Publishing.

Bone, R. C. (1991). Sepsis, the Sepsis Syndrome, Multi-Organ Failure: A Plea for Comparable Definitions. *Annals of Internal Medicine*, *114*(4), 332. https://doi.org/10.7326/0003-4819-114-4-332

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., & Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest*, *101*(6), 1644–1655. https://doi.org/10.1378/chest.101.6.1644

Brownlee, J. (2019, December 19). A Gentle Introduction to Cross-Entropy for Machine Learning. Retrieved July 11, 2020, from https://machinelearningmastery.com/cross-entropy-for-machine-learning/

Calvert, J., Saber, N., Hoffman, J., & Das, R. (2019). Machine-Learning-Based Laboratory Developed Test for the Diagnosis of Sepsis in High-Risk Patients. *Diagnostics, 9*(1), 20. doi:10.3390/diagnostics9010020

Carneiro, António Henriques, Póvoa, Pedro, & Gomes, José Andrade. (2017). Dear Sepsis-3, we are sorry to say that we don't like you. *Revista Brasileira de Terapia Intensiva*, *29*(1), 4-8. https://doi.org/10.5935/0103-507x.20170002

Centers for Disease Control and Prevention. (2020, February 14). *Data & Reports*. https://www.cdc.gov/sepsis/datareports/index.html.

Charan, R. (2020, June 19). *The Singular Value Decomposition without Algebra*. https://towardsdatascience.com/the-singular-value-decomposition-without-algebra-ae10147aab4c

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). SMOTE: Synthetic Minority Over-sampling Technique. Retrieved August 5, 2020, from https://arxiv.org/pdf/1106.1813.pdf

Ciortan, M. (2019, July 26). *Overview of feature selection methods*. Medium. https://towardsdatascience.com/overview-of-feature-selection-methods-a2d115c7a8f7.

Decision Tree Classification in Python. (2019). Retrieved July 14, 2020, from https://www.datacamp.com/community/tutorials/decision-tree-classification-python

Fermin, S. (2019, September). There is No Free Lunch in Data Science. Retrieved August 09, 2020, from https://www.kdnuggets.com/2019/09/no-free-lunch-data-science.html

Gyawali, B., Ramakrishna, K., & Dhamoon, A. S. (2019). Sepsis: The evolution in definition, pathophysiology, and management. *SAGE open medicine*, *7*, 2050312119835043. https://doi.org/10.1177/2050312119835043

Hall, M. A. (1999). Correlation-based Feature Selection for Machine Learning. Retrieved June 29, 2020, from https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Jaitley, U. (2019, April 9). *Why Data Normalization is necessary for Machine Learning models*. https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029.

Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. Cambridge, MA: The MIT Press.

Kong, Q. (2017, February 12). Machine learning 9 - More on Artificial Neural Network. Retrieved July 24, 2020, from http://qingkaikong.blogspot.com/2017/02/machine-learning-9-more-on-artificial.html

Liu, V., Escobar, G. J., Greene, J. D., Soule, J., Whippy, A., Angus, D. C., & Iwashyna, T. J. (2014). Hospital deaths in patients with sepsis from 2 independent cohorts. *JAMA*, *312*(1), 90–92. https://doi.org/10.1001/jama.2014.5804

Marik, P. E., & Taeb, A. M. (2017). SIRS, qSOFA and new sepsis definition. *Journal of thoracic disease*, *9*(4), 943–945. https://doi.org/10.21037/jtd.2017.03.125

Paoli, C. J., Reynolds, M. A., Sinha, M., Gitlin, M., & Crouser, E. (2018). Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Critical care medicine*, *46*(12), 1889–1897. https://doi.org/10.1097/CCM.0000000000003342

Remick D. G. (2007). Pathophysiology of sepsis. *The American journal of pathology*, *170*(5), 1435–1444. https://doi.org/10.2353/ajpath.2007.060872

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J. D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J. L., & Angus, D. C. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*, *315*(8), 801–810. https://doi.org/10.1001/jama.2016.0287

Sklearn.linear_model.LogisticRegression. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Sklearn.linear_model.LogisticRegression. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Sklearn.naive_bayes. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/classes.html#module-sklearn.naive_bayes

Sklearn.svm.LinearSVC. (2020). Retrieved July 09, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html

Sklearn.tree.DecisionTreeClassifier. (2020). Retrieved August 09, 2020, from https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Sprung, C. L., Sakr, Y., Vincent, J.-L., Gall, J.-R. L., Reinhart, K., Ranieri, V. M., … Payen, D. (2006). An evaluation of systemic inflammatory response syndrome signs in the Sepsis Occurrence in Acutely ill Patients (SOAP) study. *Intensive Care Medicine*, *32*(3), 421–427. https://doi.org/10.1007/s00134-005-0039-8

Support vector machines. (n.d.). Retrieved July 26, 2020, from http://statweb.stanford.edu/~jtaylo/courses/stats202/svms.html

Tong, L., Luo, J., Cisler, R., & Cantor, M. (2019). Machine Learning-Based Modeling of Big Clinical Trials Data for Adverse Outcome Prediction: A Case Study of Death Events. *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*. doi:10.1109/compsac.2019.10218

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., . . . Zhao, S. (2019, April 11). Applications of machine learning in drug discovery and development. Retrieved August 1, 2020, from https://www.nature.com/articles/s41573-019-0024-5

Vert, J., Tsuda, K., & Scholkopf, B. (2004). A Primer on Kernel Methods. Retrieved July 7, 2020, from http://members.cbio.mines-paristech.fr/~jvert/publi/04kmcbbook/kernelprimer.pdf

Vincent J. L. (2016). The Clinical Challenge of Sepsis Identification and Monitoring. *PLoS medicine*, *13*(5), e1002022. https://doi.org/10.1371/journal.pmed.1002022

Vincent, J. L., Opal, S. M., Marshall, J. C., & Tracey, K. J. (2013). Sepsis definitions: time for change. *Lancet (London, England)*, *381*(9868), 774–775. https://doi.org/10.1016/S0140-6736(12)61815-7

Yiu, T. (2019, August 14). Understanding Random Forest. Retrieved August 1, 2020, from https://towardsdatascience.com/understanding-random-forest-58381e0602d2