Systems Engineering-assisted Machine Learning for Biomedical Applications

by

Farnaz Yousefi Zowj

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
May 4, 2024

Keywords: Feature Engineering, Machine Learning, Speech Disorder, Autism Spectrum Disorder, Pulmonary Arterial Hypertension

Approved by

Peter He (chair), George E. & Dorothy Stafford Uthlaut Endowed Professor of Department of Chemical Engineering, Auburn University
Jin Wang, Walt and Virginia Woltosz Professor of Department of Chemical Engineering, Auburn University
Selen Cremaschi, B. Redd Professor of Department of Chemical Engineering, Auburn University
Marisha Speights Atkins, Assistant Professor of Communication Sciences & Disorders, Northwestern University
Amit Kumar Mitra, Assistant Professor of Harrison College of Pharmacy, Auburn University

Abstract

In recent years, advancements in biomedical technology have led to the accumulation of vast amounts of healthcare data. Machine learning (ML) algorithms are now employed to extract valuable insights from data. However, the challenge lies in extracting meaningful information from big data due to irrelevant data and noise. Integrating domain knowledge into ML techniques is crucial to address these challenges and deliver comprehensive and interpretable results.

Speech disorders in children pose diagnostic challenges due to intra- and inter-rater variabilities in auditory perceptual analysis (APA) and manual transcription methods. To overcome these limitations, we explore the utilization of Landmark (LM) analysis with novel knowledge-based features for automatic speech disorder detection. Our systematic study shows nearly a 20% improvement in accuracy, highlighting the effectiveness of these features in classifying speech disorder patients.

A robust framework is proposed that integrates ML techniques with domain knowledge for detecting autism spectrum disorder (ASD) using serum biomarkers. Despite challenges in identifying reliable biomarkers due to protein level variations, our framework outperforms previous methods by integrating feature engineering and selection with linear ML algorithms, achieving high performance in ASD detection. The proposed framework improves the area under the curve (AUC) by 10%, demonstrating its effectiveness in reducing within-class variations.

Furthermore, we investigate the early detection of pulmonary arterial hypertension (PAH) in systemic sclerosis (SSc) patients using proteomic data. Many ML-assisted detection frameworks are limited to the dataset used for training, and they may not perform effectively when applied to different diseases or disorders. This case study underscores the ASD detection framework's effectiveness for detecting various diseases and disorders. The proposed framework achieves over 16% enhancement in PAH detection accuracy from previous detection models.

Our study highlights the efficacy of combining ML with domain knowledge for disorder detection. The feature engineering and selection techniques enhance the robustness and reliability of early detection of disorders, emphasizing the importance of knowledge-guided models for interpretable results.

Acknowledgments

I would like to extend my heartfelt gratitude to my advisor, Dr. Peter He, for his unconditional support and guidance throughout my graduate studies. His mentorship shaped my academic growth.

A special acknowledgment goes to Dr. Jin Wang for her belief in me and continuous support. I would like to thank the rest of my committee members, Dr. Selen Cremaschi, Dr. Marisha Speights Atkins, and Dr. Amit Kumar Mitra for their thorough review of my work and invaluable feedback.

I would like to thank both current and past members of the He lab, including Kerul, Jisung, and Alex. Kerul, in particular, consistently provided assistance, even amidst his busy schedule, for which I am truly thankful.

I am deeply appreciative of the friendship and companionship provided by Bahareh, Mehdi, and Mehran, which their presence has made my time at Auburn truly memorable. I would also like to thank my dear friend Fazile for her words of encouragement over the years.

I am profoundly grateful to my family for their unconditional love and support, specially Parisa, and Afsoon whose presence has brought warmth and comfort to my life in the United States.

Last but not least, I would like to express my deepest appreciation to my husband, Amir, whose unconditional support, encouragement, and positivity have been constant sources of inspiration. He is truly my ray of sunshine.

I dedicate this dissertation to my husband who believed in me during my moments of doubt.

Table of Contents

Abstract ii

Acknowledgments iv

List of Abbreviations xii

1 Introduction 1

2 Feature Engineering and Machine Learning for Computer-assisted Screening of Children with Speech Disorders 3

2.1 Introduction 3

2.2 Materials and methods 7

2.2.1 Ethical considerations 7

2.2.2 Speakers 7

2.2.3 Dataset 8

2.2.4 Feature engineering 8

2.2.5 Feature selection 8

2.2.6 Sample imbalance 10

2.2.7 Synthetic minority over-sampling technique 11

2.2.8 Monte-Carlo cross validation and testing 12

2.2.9 The trade-off between sensitivity and specificity 14

2.2.10 Classification techniques 14

2.2.10.1 Linear discriminant analysis 15

2.2.10.2 Support vector machine 16

2.2.10.3 Random forest . . . 17
2.2.10.4 Extreme gradient boosting . . . 17
2.2.11 Effect of adding triage words . . . 17
2.3 Results and discussion . . . 18
2.4 Conclusion . . . 26

3 A Systems Engineering Computer-assisted Biomarker Detection Framework for Autism Spectrum Disorder using Proteomic Data . . . 28
3.1 Introduction . . . 29
3.2 Previous works . . . 32
3.3 Materials and Methods . . . 34
3.3.1 Ethical considerations . . . 34
3.3.2 Participants . . . 34
3.3.3 Dataset . . . 34
3.3.4 Monte-Carlo cross validation and testing . . . 35
3.3.5 Feature engineering . . . 35
3.3.6 Data preprocessing . . . 36
3.3.7 Feature selection . . . 37
3.3.7.1 Filter feature selection technique . . . 37
3.3.7.2 Wrapper feature selection technique . . . 38
3.3.7.3 Hybrid feature selection technique . . . 39
3.3.8 The feature selection algorithm . . . 40
3.3.8.1 Feature pre-selection with Fisher's criterion . . . 40
3.3.8.2 The filtering criterion . . . 41
3.3.8.3 Sequential feature selection . . . 45
3.3.8.4 Final panel of biomarkers . . . 47
3.3.9 Classification techniques . . . 49

3.3.9.1 Linear discriminant analysis . . . 49
3.3.9.2 Logistic regression . . . 50
3.3.9.3 Support vector machine . . . 51
3.3.9.4 Gaussian Naïve Bayes . . . 52
3.3.10 Receiver operator characteristic curve . . . 53
3.4 Results . . . 55
3.4.1 Feature engineering . . . 55
3.4.2 Feature selection . . . 55
3.4.2.1 Pre-selection based on Fisher's criterion . . . 55
3.4.2.2 Hybrid feature selection . . . 58
3.4.3 Performance evaluation . . . 60
3.5 Discussion . . . 61
3.6 Conclusion . . . 70

4 Detecting Pulmonary Arterial Hypertension: A Case Study Using the Biomarker Detection Framework from chapter 3 . . . 71
4.1 Introduction . . . 72
4.2 Materials and Methods . . . 73
4.2.1 Dataset . . . 73
4.2.2 Preprocessing . . . 73
4.2.3 Methods . . . 74
4.3 Results . . . 74
4.4 Discussion . . . 74
4.5 Conclusion . . . 78

5 Summary and Future Work . . . 80

References . . . 95

List of Tables

2.1 Description of LMs used in this study 4

2.2 Features employed in this study 9

2.3 The ten features selected based on RFECV 10

2.4 LMs detected for a person uttering 11 triage words 19

2.5 LMs detected for a person uttering 11 triage words put together in a sentence format 20

2.6 Comparison of classification performance based on raw features 20

2.7 Comparison of classification performance based on rationally engineered and selected features 22

2.8 LDA classification performance when different class priors are used 24

3.1 Selected features for training data and their Fisher's scores 56

3.2 Fisher's scores for the constituent proteins of the selected ratios 57

3.3 Fisher's scores for the top 10 ratios with the highest scores 57

3.4 Fisher's scores for the top 10 proteins with the highest scores 57

3.5 Fisher's scores for the 9 proteins suggested in [43] 58

3.6 Point biserial correlation coefficients of the selected ratios with label 58

3.7 Point biserial correlation coefficients of the constituent proteins of the selected ratios with label 58

3.8 Point biserial correlation coefficients of the 9 proteins suggested in [43] with label 59

3.9 AUC of LR, LDA, SVM, and GNB classifiers on the panel of 9 biomarkers proposed by [43] 61

3.10 AUC of LR, LDA, SVM, and GNB classifiers on the panel of 8 biomarkers proposed by our algorithm 62

3.11 AUC of the algorithm with different parameters 62

4.1 AUC of LR, LDA, SVM, and GNB classifiers on the panel of 8 biomarkers proposed by our algorithm 75

4.2 AUC of LR, LDA, SVM, and GNB classifiers on the panel of 6 proteins proposed by [118] 76

List of Figures

2.1 Detected LMs using SpeechMark®. 5

2.2 RFECV 10

2.3 Schematic of the MCVT for comparing different modeling techniques and assessing their performances in terms of accuracy and robustness. 13

2.4 The trade-off between sensitivity and specificity illustrated by a ROC curve. . . 15

2.5 Comparison of classification performance based on raw features. 21

2.6 Comparison of classification performance when selected features are used. . . . 22

2.7 The impact of class priors on sensitivity, specificity and accuracy of the LDA classifier. 23

2.8 ROC curve of the LDA classifier. 25

3.1 Schematic of the MCVT procedure. 36

3.2 The hybrid FS algorithm applied in each MCVT. 47

3.3 The ROC curve illustrating the trade-off between false positive and negative rates. 54

3.4 The plot of average AUC for a range of number of features. The selection is based on the AUC of validation (red plot). The global maximum of this curve is shown with a green dotted line. The optimal model size is shown with a yellow dotted line. 56

3.5 The ROC curve of the 8 suggested features using LDA, LR, SVM, and GNB classifiers. 63

3.6 The boxplot of the 9 proteins suggested by [43]. The median line is colored in green, whisker lines in purple, and the minimum and maximum lines in yellow. The cases are colored blue while the controls are orange. The outliers are colored in black. 64

3.7 The boxplot of the 8 features suggested by our algorithm. The median line is colored in green, whisker lines in purple, and the minimum and maximum lines in yellow. The cases are colored blue while the controls are orange. The outliers are colored in black. 65

4.1 The plot of average AUC for a range of number of features. The selection is based on the AUC of validation (red plot). The global maximum of this curve is shown with a green dotted line. The optimal model size is shown with yellow dotted line. 75

List of Abbreviations

ABAS-II Adaptive Behavior Assessment System-Second Edition

ADI-R Autism Diagnostic Interview–Revised

ADOS Autism Diagnostic Observation Schedule

Ang-2 Angiopoietin-2

APA Auditory Perceptual Analysis

ASD Autism Spectrum Disorder

ASDD Autism Spectrum Disorder Discovery

ASR Automatic Speech Recognition

AUC Area Under the Curve

b2-Microglobulin Beta-2-Microglobulin

BMI Body-Mass Index

C5b, 6 Complex Complement C5b-C6 Complex

CART Classification And Regression Trees

CFH Complement Factor H

CFS Correlation-Based Feature Selection

CV Cross-Validation

DERM Dermatopontin

DL Deep Learning

DPP2 Dipeptidyl Peptidase 2

DR6 Death Receptor 6

DT Decision Tree

EGFR Epidermal Growth Factor Receptor

ErbB3 Receptor Tyrosine-Protein Kinase Erbb-3

ESKD End-Stage Kidney Disease

ET Ensemble Technique

FAM3D Family With Sequence Similarity 3 Member D

FDA Fisher Discriminant Analysis

FS Feature Selection

GNB Gaussian Naïve Bayes

GRN Gene Regulatory Network

HDAC1 Histone Deacetylases 1

HLP High Leverage Point

hnRNP K Heterogeneous Nuclear Ribonucleoprotein K

HRT Hybrid Rejection Technique

IgD Immunoglobulin D

IGF Insulin-Like Growth Factor

IL-1 Interleukin-1

IL-1RII Interleukin-1 Receptor Type 2

IL-1Rrp2 Interleukin-1 Receptor-Like 2

IL-6 Interleukin-6

IL-6 SRa Interleukin-6 Receptor Subunit Alpha

IL-6R beta Interleukin-6 Receptor Subunit Beta

IPAH Idiopathic Pulmonary Arterial Hypertension

KLK-7 Kallikrein-7

KNN K-Nearest Neighbor

LDA Linear Discriminant Analysis

LM Landmark

LR Logistic Regression

lsqr Least Square

M-CSF R Macrophage Colony-Stimulating Factor 1 Receptor

MAC Membrane Attack Complex

MAPK6 Mitogen-Activated Protein Kinase 6

MCVT Monte-Carlo Cross Validation and Testing

MHC Major Histocompatibility Complex

MIF Migration Inhibitory Factor

ML Machine Learning

mPAP Mean Pulmonary Arterial Pressure

MS Mass Spectrometry

NB Naïve Bayes

NT-proBNP N-Terminal Pro-Brain Natriuretic Peptide

P-Cadherin Cadherin-3

PAH Pulmonary Arterial Hypertension

PAI-1 Plasminogen Activator Inhibitor-1

PCA Principal Component Analysis

PCWP Pulmonary Capillary Wedge Pressure

PH Pulmonary Hypertension

PRS Precise Rejection Stage

QRS Quick Rejection Stage

RBF Radial Basis Function

RCA Reversed Correlation Algorithm

RF Random Forest

RFE Recursive Feature Elimination

RFECV Recursive Feature Elimination with Cross-Validation

RHC Right Heart Catheterization

ROC Receiver Operator Characteristic

SBFS Sequential Backward Floating Search

SBS Sequential Backward Selection

SC Syllabic Cluster

SEED Speech Evaluation and Exemplars Database

SFFS Sequential Forward Floating Search

SFS Sequential Forward Selection

SMOTE Synthetic Minority Over-Sampling Technique

SRCN1 SRC Kinase Signaling Inhibitor 1

SSc Systemic sclerosis

SuHx Sugen 5416/Chronic Hypoxia

svd Singular Value Decomposition

SVM Support Vector Machine

TD Typically Developing

TIMP-1 Tissue Inhibitor of Metalloproteinases 1

TNF Tumor Necrosis Factor

VEGF-D Vascular Endothelial Growth Factor D

XGBoost Extreme Gradient Boosting

Chapter 1

Introduction

In recent years, with advancements in biomedical technology and healthcare data management systems, industries have been able to accumulate vast amounts of data and leverage ML and deep learning (DL) algorithms to extract valuable insights. Despite the potential of ML approaches, extracting meaningful information from big data can be challenging due to irrelevant data and noise, and the results of pure data-driven ML techniques can sometimes lead to misleading conclusions. This highlights the importance of integrating domain knowledge into ML techniques to fill in their gaps.

Systems engineering involves the analysis, development, management, and evaluation of complex systems to improve their efficiency, reliability, and safety. This approach transforms traditional data-driven models by integrating ML with human knowledge (domain knowledge). Domain knowledge, which pertains to the specific field of the data, proves especially beneficial in feature engineering, where features are tailored to enhance ML algorithms. By incorporating domain knowledge into ML models, this study aims to deliver comprehensive and interpretable results and enhance decision-making processes through effective data-driven solutions.

In this research, across three chapters, I present frameworks developed to detect speech disorders, ASD, and PAH in SSc patients.

In the first part of this work (Chapter 2), I introduce a novel speech disorder detection framework by integrating feature engineering and feature selection (FS) techniques into ML models. APA and manual transcription methods face challenges in diagnosing speech and language deficits in children, including intra- and inter-rater variabilities. To address these limitations, we explore the utilization of LM analysis coupled with novel knowledge-based

features such as ratio- and strength-based features for automatic speech disorder detection. A systematic study and comparison of different linear and nonlinear ML classification techniques based on the raw features and the proposed features is conducted to assess the effectiveness of the novel features in classifying speech disorder patients from normal speakers. Results suggest nearly 20% improvement in terms of accuracy.

In the second part of this dissertation (Chapter 3, I propose a robust framework integrating ML techniques with domain knowledge for detecting ASD using serum biomarkers. While behavioral criteria are used as the standard for ASD diagnosis, recent proteomic analyses show metabolic differences in the plasma/serum of individuals with ASD. Yet identifying reliable biomarkers remains challenging due to significant variations in protein levels caused by confounding factors. To tackle this challenge, I introduce an automated biomarker detection framework, which integrates novel ratio-based features with a hybrid FS method and a linear ML model. Our ASD detection framework outperforms previous methods by 10% in terms of the AUC due to the novel features, while reducing within-class variations.

In the third section (Chapter 4), I investigate the early detection of PAH in SSc patients with proteomic data, highlighting the efficacy of combining ML with domain knowledge within the same framework developed for ASD. This case study underscores the framework's effectiveness for detecting various diseases and disorders. The feature engineering and FS techniques presented in this work enhance the robustness and reliability of early detection of the disorders proving the importance of knowledge-guided models for more interpretable results. The proposed framework can detect PAH with more than 16% improvement in terms of AUC.

The conclusion and potential future work are summarized in the final chapter of this dissertation (Chapter 5). This study underscores the limitations of pure data-driven ML methods, particularly their unreliable results, and emphasizes the importance of integrating domain knowledge into model development to establish more comprehensive and interpretable models. Novel techniques for feature engineering and selection are proposed to enhance model robustness and reliability and reduce the effect of noise.

Chapter 2

Feature Engineering and Machine Learning for Computer-assisted Screening of Children with Speech Disorders

Auditory perceptual analysis (APA) is the main method for clinical assessment of speech-language deficits, which are one of the most prevalent childhood disabilities. However, results from APA are susceptible to intra- and inter-rater variabilities. There are also other limitations of manual or hand transcription-based speech disorder diagnostic methods. There is increased interest in developing automated methods that quantify speech patterns for diagnosing speech disorders in children to address these limitations. LM analysis is an approach that characterizes acoustic events occurring due to sufficiently precise articulatory movements. This work investigates the utilization of LMs for automatic speech disorder detection in children. Besides the LM-based features that have been proposed in existing research, we propose a set of novel knowledge-based features that have not been proposed before. A systematic study and comparison of different linear and nonlinear ML classification techniques based on the raw features and the proposed features is conducted to assess the effectiveness of the novel features in classifying speech disorder patients from normal speakers.

2.1 Introduction

Speech-language deficits are one of the most prevalent childhood disabilities affecting about 1 in 12 children between three and five years old [1]. Despite the recognition that early identification and treatment of communication disorders is important for school readiness and has been shown to significantly improve communication, literacy, and mental health outcomes for young children [1, 2, 3], approximately 40% of children with speech and language disorders do not receive intervention because their impairment goes undetected [4, 5]. APA is the main

method for clinical assessment of disordered speech; however, results from APA are susceptible to intra- and inter-rater variabilities [6]. Another factor to consider is that some children may be reluctant to participate in long testing sessions [7], and even if they do, transcription of large data sets of audio recordings is time-consuming and requires a high level of expertise from therapists [8, 9]. These limitations of manual or hand transcription based diagnostic assessment methods have led to an increasing need for auto- mated methods to quickly and consistently quantify child speech patterns and help them be diagnosed if they have impaired speech [10]. LM analysis is such an approach that characterizes speech with acoustic markers that are developed based on the LM theory of speech perception [11, 12, 13]. Unlike automatic speech recognition (ASR), LM analysis does not attempt to identify words, but rather to detect acoustic events that occur as the result of sufficiently precise articulatory movements. LM analysis has been suggested as the basis for automatic speech analysis [6]. Therefore, in this work we focus on the utilization of LMs for automatic speech disorder detection in children, with LMs extracted using a publicly available software: SpeechMark toolbox [6]. SpeechMark not only analyzes physical aspects of the signal but also applies acoustic knowledge of articulatory features in the process of analysis [6]. As a result, it has been utilized in numerous studies to extract LMs for various applications such as the detection of stress [14], depression [15], emotion [16], and sleep deprivation [17]. Here we briefly review how it works. SpeechMark divides the computed spectrogram into six frequency bands, and then fine and coarse processing steps are conducted to detect band-energy rise and to determine threshold for peak detection. Finally, energy peaks are located and LM types are determined based on the patterns of changes in the frequency bands [12, 13]. The description of each LM detected by this tool and used in this study are presented in Table 2.1. An example of LMs detected by SpeechMark from the speech of a speaker uttering a word are shown are shown in Figure 2.1.

Table 2.1: Description of LMs used in this study

LM	Description
g (glottis)	Onset (+) and offset (-) of sustained motion of vocal fold
b (burst)	Onset (+) and offset (-) of frication or bursts in an unvoiced segment
s (syllabicity)	Release (+) and closure (-) of sonorant consonant in voiced segment
f (unvoiced frication)	Onset (+) and offset (-) of frication in an unvoiced segment
v (voiced frication)	Onset (+) and offset (-) of frication of in a voiced segment

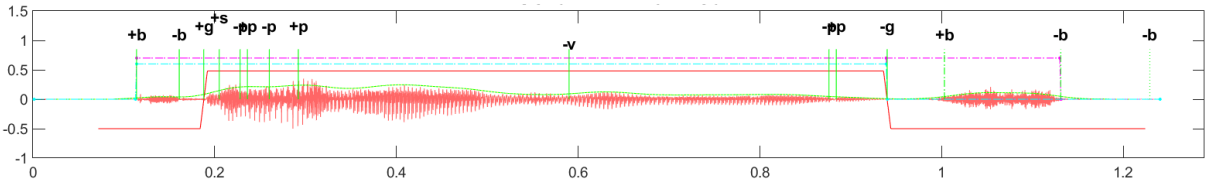

Figure 2.1: Detected LMs using SpeechMark®.

An extension of SpeechMark for the LM system, namely automatic syllabic cluster (SC) analysis, was recently proposed that clusters the LMs into syllabic units [18]. These LMs are grouped based on specific rules, such as at least a 30 ms voiced segment is required in a SC [19]. Studies show that SC patterns are good indicators of differences between normal and disordered speakers [18, 19]. Variations in articulatory exactness in normal and disordered speakers have proven to be related to LM and SC patterns [19]. This is validated by existing studies showing that simple count of LMs and/or SCs can be used for classification of gender [6], Parkinson's disease [20], and sleep deprivation [17]. Counting of individual LMs, a.k.a. unigrams, does not consider the specific order or sequence of the LMs, which may contain important information about the speech. n-gram, which is a generalization of unigrams and is defined as a sequence of n consecutive LMs, takes the specific LM order into consideration when $n \geq 2$ [21]. It was found that n-gram counts (n = 1,2,3,4) were good features for depression detection [15, 22]. For example, in [22], SpeechMark was used to extract LMs from a large dataset consisting of recordings from smartphones. Two sets of features were proposed based on speech LM bigrams, i.e., bigram-count and LDA-bigram. The first set calculates the frequencies of bigrams, and the second set detects latent patterns from bigrams using natural language text processing. A linear support vector machine (SVM) classifier was trained using the two sets of features. It was found that the bigram features increased the accuracy of the SVM classifier from 72.9% when only acoustic features were used to 78.7% when either bigram-count or LDA-bigrams were utilized. The speech LM bigram features improved the F1(depressed) by 30.1% compared to acoustic features [22]. Besides n-gram count, time-based LM features have also been proposed in the literature. These time-based LM features include durations of the bigrams (i.e., 2-grams) and LM pairs (i.e., onset and offset of a LM as defined

in Table 2.1) [15], and speech rate, which is defined as the number of phonetic units, such as syllables or words, uttered per unit time [23, 24].

In this work, we have adopted count of n-grams as well as duration and rate features based on LMs and n-grams. One contribution of this work is to propose novel knowledge-based features that have not been proposed before and to demonstrate the effectiveness of these new features in detecting childhood speech disorder. For example, this work studies features that are the ratios of the count of n-grams ($n \geq 2$) to that of unigrams. The idea of considering ratio is similar to the body-mass index (BMI) where the weight itself cannot determine whether a per- son is overweight or not. BMI takes the height of the person into account as well. The ratios are usually better features than the absolute individual values in addressing the individual variations of samples within the same class. This point is further validated in this study. Another contribution is to perform systematic FS to identify key features and quantify their contributions to the classification of patients with speech disorder from normal controls. The final contribution of this work is a systematic study and comparison of different linear and nonlinear ML classification techniques and their effectiveness in classifying speech disorder patients from normal speakers.

The remainder of this work is organized as follows. Section 2.2 describes materials used in this study, which include the general information about the speakers of which the speech samples were collected, the conditions and procedures the speech samples were processed to obtain the dataset, and the features proposed to be studied in this work. Section 2.2 also introduces the analytical methods used in this study, which include methods to address the data imbalance, introduction of ML classification techniques used in this work, and the procedure and criteria used to evaluate the performance of different ML techniques in screening children with speech disorders. Section 2.3 presents results and discussions of this work, and Section 2.4 draws some conclusions.

2.2 Materials and methods

2.2.1 Ethical considerations

Ethical approval for human subject data collection was granted by the University of Cincinnati for this study with reference number 2015–3023 and subsequently at Auburn University with reference number 17–203 EP 1705 for ongoing data analysis and additional data collection. Permissions were also sought from schools and university clinics where data were collected. Anonymity and confidentiality were explained to participants. Participants were assured that withdrawal from study would not harm them in any way. Informed consent forms were filled and signed by parents with verbal assent from the child participants. Participant's data was de-identified with codes to ensure anonymity. All data analyses in this work are conducted using the de-identified data.

2.2.2 Speakers

The speech of 52 children ages 33–94 months (with mean 51.52 and standard deviation 10.16) was retrieved from the Speech Evaluation and Exemplars Database (SEED) [25]. Due to missing values, one sample was dropped from this work. Of the 51 remaining children, 39 were typically developing (TD) without speech or language disorder, and 12 were diagnosed with speech sound disorder without language impairment. All children were required to demonstrate normal hearing using the criterion of sound detection at 20 dB HL for pure tones at 500, 1000, 2000, and 4000 Hz. Participants were required to exhibit age-appropriate receptive language skills on the CELF Preschool-2 [26]. Age-appropriate performance was determined by scores falling within one standard deviation of the mean (standard score > 85). Children were classified as TD or with speech disorder using the Clinical Assessment of Articulation and Phonology-2 or the Diagnostic Evaluation of Articulation [27]. Children with standard scores ≤ 85 (one standard deviation below the mean) were assigned to the with speech disorder group. Children with concomitant language disorders were not included in the study.

2.2.3 Dataset

The speech samples retrieved were recorded in local community early education centers or in the lab. Sound levels were measured prior to each recording session to determine if the environmental noise level was below 40 dBA SPL in both the school and lab environment [28]. Speech samples were recorded at a 44K sampling rate at 24-bit depth using a handheld ZOOM H6N recorder (Zoom North America) with cardioid XLR MOVO LV402 microphones (MOVO). Speech samples retrieved for this study were one of the Triage 11 word set from Anderson and Cohen [29]: *flower*. Acoustic LMs, including +/-g, +/-b, +/-s, +/-f, and +/-v, as well as SCs were obtained using the Speech- Mark MATLAB toolbox (STAR Corp., MA).

2.2.4 Feature engineering

The raw features extracted from audio recordings using the SpeechMark Toolbox include time stamp and strength of each LM listed in Table 2.1, plus SC count. As discussed previously in Section 2.1, in this work we have adopted all LM and SC based features proposed in the literature, including n-gram counts, and duration and rate features based on LMs and n-grams. These features are listed in the top rows of Table 2.2. In addition, we explore LM strength based features and propose n-gram ratio based features to better address within-class variations as discussed in Section 2.1. These new features are listed in the bottom rows of Table 2.2. After removing illegitimate or trivial features (e.g., n-gram counts that are all zeros, or ratios with a denominator of zero), there are 303 unique features generated based on the criteria listed in Table 2.2.

2.2.5 Feature selection

It has been shown by many studies that the performances of classification methods can be significantly improved if only the relevant features are included as the predictors. FS can also reduce the risk of overfitting, which is especially important when the number of samples are relatively small compared to the number of features (such as the case of this study). Finally, FS can reduce model complexity, making result interpretation easier. As a result, FS has been one of the most important practical concerns in data-driven approaches. In the past few decades,

Table 2.2: Features employed in this study

Feature category	Description	Unit
Features adopted from literature		
Unigram count	Number of each unigram type	#
Bigram count	Number of each bigram type	#
Trigram count	Number of each trigram type	#
Average bigram duration	Average duration of all bigrams	s
Average trigram duration	Average duration of all trigrams	s
Duration of LM pair	Average duration of LM pairs of each LM type	s
Unigram rate	Count of all unigram types per unit time	#/s
Bigram rate	Count of all bigram types per unit time	#/s
Trigram rate	Count of all trigram types per unit time	#/s
SC count	Number of SCs	#
Speech rate	SC count per unit time	#/s
New features proposed in this work		
Strength of unigram	Average strength of each unigram type	%
Strength of bigram	Average strength of each bigram type	%
Strength of trigram	Average strength of each trigram type	%
Strength change	Average strength difference of two consecutive LMs	%
Average bigram strength	Average strength of all bigram types	%
Average trigram strength	Average strength of all trigram types	%
Unigram/unigram ratio	Ratio of unigram counts of each type	-
Bigram/unigram ratio	Ratio of bigram count to unigram count of each type	-
Trigram/unigram ratio	Ratio of trigram count to unigram count of each type	-

many different FS approaches have been reported for various modeling and classification applications. For more detailed discussions on various FS methods, the readers are referred to some recent review articles.

In this work, a two-step FS procedure is proposed. In the first step, the redundant features (i.e., the features that are highly correlated with an existing feature) are removed. In the study, a Pearson correlation coefficient of 0.99 is used as the criterion to determine whether a feature is redundant with an existing feature or not. After this step, the number of features is reduced to 189 from the original 303 features, indicating that there is significant redundancy among the original features.

In the second step, the recursive feature elimination with cross-validation (RFECV) from *scikit-learn* is utilized with the default 5-fold cross-validation (CV). Figure 2.2 shows the CV score vs. number of features when a linear discriminant analysis (LDA) model is used as the classifier. Figure 2.2 indicates that only 10 features are needed to obtain the optimal CV score.

The 10 features selected are listed in Table 2.3. As can be seen from Table 2.3, nine out of the ten features are new features proposed in this work that have not been utilized before. Among the nine new features, seven are ratio-based features and two are strength-based features.

Table 2.3: The ten features selected based on RFECV

Feature category	**Feature specifics**
Ratios of bigram count to unigram count	'-g-b/+g'
Ratios of trigram count to unigram count	'-s+s-s/+g'
Ratios of trigram count to unigram count	'+b+g+s/+g'
Ratios of bigram count to unigram count	'-b+g/+g'
Trigram counts	'+b-b+b'
Ratios of trigram count to unigram count	'-s-g+b/+g'
Ratios of trigram count to unigram count	'-g+b-b/+g'
Strength of trigrams	'+g-g-b'
Strength of unigrams	'-f'
Ratios of trigram count to unigram count	'+b+g-v/+g'

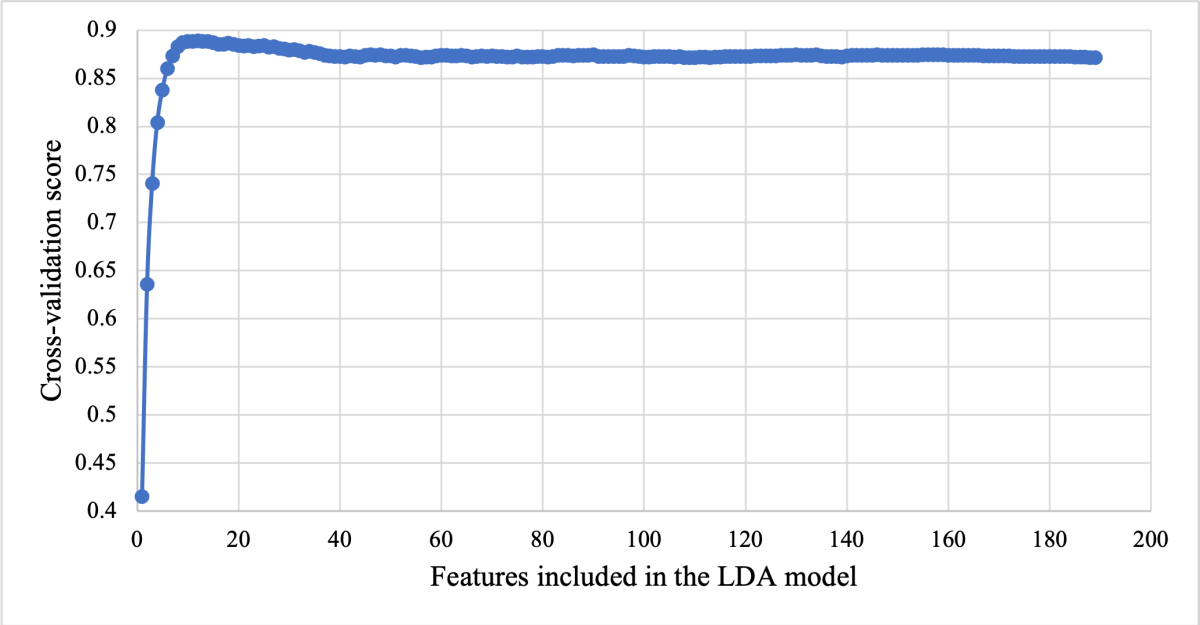

Figure 2.2: RFECV

2.2.6 Sample imbalance

Among all 51 samples, 39 samples belong to normal speakers, while the remaining 12 samples belong to the disordered speakers, which indicates an approximately 3:1 class imbalance between the normal speaker samples and the disordered ones. However, most ML classification algorithms are developed with the implicit assumption of approximately equal samples

in each class. Therefore, data with imbalanced or skewed classes may result in poor classification performance for the minority class samples. For example, if the model is tuned using accuracy, the resulted model may lead to mostly correct classification of the majority class at the cost of poor classification of the minority class. However, the correct classification of minority class samples is often more critical as they represent the disease group most of the time– misclassification of these samples leads to low sensitivity. Several ways of dealing with class imbalance have been proposed in the literature such as under-sampling, over-sampling, synthetic sample generation, using cost-sensitive methods, and applying penalties or weights based on class ratio [30]. Under-sampling reduces the number of samples in the majority class to improve the imbalance ratio, while oversampling refers to increasing the number of samples in the minority class samples. Oversampling is used more often than under-sampling to maximally utilize available samples. Random oversampling refers to an increase in minority class samples through duplication of randomly selected minority class samples. However, this most straightforward approach that duplicates the existing samples does not add new information during training and is not considered robust. A more robust method when oversampling a dataset is the synthetic minority over-sampling technique (SMOTE), in which new samples are synthesized from the existing samples [31]. SMOTE is utilized in this work to address the class imbalance issue and we briefly review the technique and the implementation details in the following subsection.

2.2.7 Synthetic minority over-sampling technique

Based on the feature space of the minority samples, SMOTE first selects a minority class instance or sample at random (denoted as a) and then finds its k nearest minority class neighbors. The synthetic neighbor is created by selecting one of the k nearest neighbors at random (denoted as b) and connecting them to form a line segment in the feature space. The synthetic sample is generated as a combination or linear interpolation between the two chosen samples, a and b, as follows.

$$x_{new} = x_a + \lambda(x_b - x_a). \tag{2.1}$$

where x_i denotes the feature vector (i.e., a point in the feature space) of sample i, λ is a random number in the range [0, 1]. For features that only take integer values (e.g., n-gram counts), x_{new} is rounded to the nearest integer. More information on SMOTE can be found in [31]. For implementation, Python-based library *imbalanced-learn* was used in our work for SMOTE oversampling [32]. Due to limited data, we first randomly isolate one sample from each class for testing. Once the two random samples, i.e., one from a normal speaker and one from a disordered speaker, are removed from the set, we apply SMOTE oversampling to balance the dataset. The primary purpose behind separating test samples before oversampling is to avoid bias in the model due to the test samples' influence on the synthetic samples created. After removing one sample from each class for testing, we have 38 samples from normal speakers and 11 samples from disordered speakers in the training set. SMOTE oversampling is applied on the training set, where 27 samples in the minority class, i.e., disordered speakers, are generated to balance the dataset.

2.2.8 Monte-Carlo cross validation and testing

Once the training set is balanced, we train different classification models, perform FS and tune their hyperparameters using 10-fold CV on the training set. Then we apply the models to the left-out test samples, and report the sensitivity and specificity of each model. This whole procedure is referred to as one Monte-Carlo cross validation and testing (MCVT) [33]. We report the mean and standard deviation of sensitivity and specificity of 50 such MCVT runs, which is a robust way of comparing different modeling techniques and assessing their performances. MCVT avoids overfitting by randomly selecting and isolating two test samples first, then utilizing SMOTE technique to balance the rest of the dataset, which is further split into training and validation for modeling training, FS, and hyperparameter tuning. For hyperparameter tuning, we use a 10-fold stratified CV to select the optimal hyperparameters. The schematic of the proposed MCVT procedure is shown in Figure 2.3.

Sensitivity and specificity are two most commonly used critical metrics when dealing with binary classification problems in healthcare. Sensitivity is the true positive rate, i.e., the classifier's ability to detect diseased patients correctly, and specificity is the true negative rate,

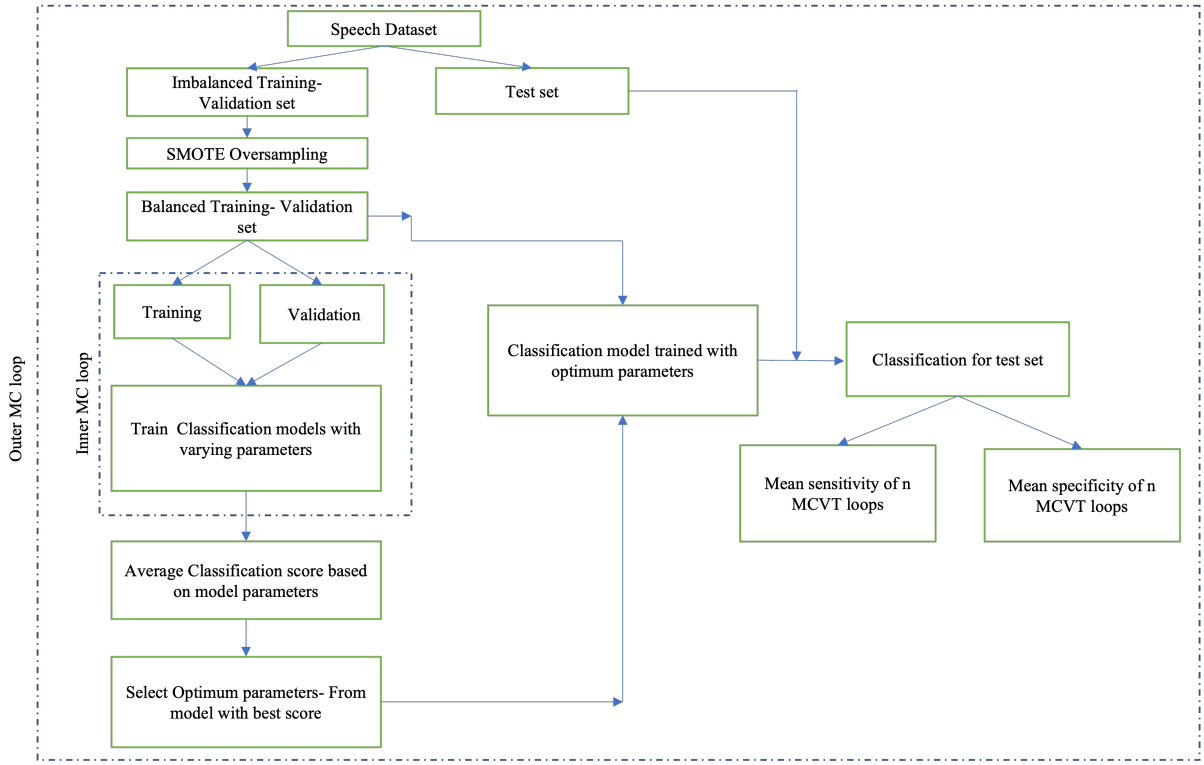

Figure 2.3: Schematic of the MCVT for comparing different modeling techniques and assessing their performances in terms of accuracy and robustness.

i.e., the classifier's ability to detect normal controls (i.e., the ones without diseases) correctly. We also use accuracy as a single measure when we need to evaluate the overall performance of a classifier. The mathematical definitions of these terms are given below.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

$$Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \tag{2.4}$$

where n_{TP} is the number of true positives, n_{FN} number of false negatives, n_{TN} number of true negatives, and n_{FP} number of false negatives. Sensitivity, specificity and accuracy all range from 0 to 1 (or 0 to 100%).

As shown in Figure 2.3, the mean or average sensitivity and specificity of MCVT runs can be used to assess the accuracy of a classifier; while the standard deviations of the sensitivity and specificity of the MCVT runs can be used to quantify the robustness of a classifier (i.e., how

consistently a classifier performs when trained with randomly selected training samples). It is worth noting that because only one sample from each class is left out for testing in this work, the standard deviations of the sensitivity and specificity would be biased due to the extremely small sample size. Therefore, this measure of robustness is not utilized in this work.

2.2.9 The trade-off between sensitivity and specificity

For binary classification, there is often a trade-off between sensitivity and specificity. This trade-off can be visualized in a receiver operator characteristic (ROC) curve, which plots sensitivity vs. (1 –specificity) as illustrated in Figure 3.5. The AUC is the summative measure of the classification capability of a classifier. A perfect classifier has an AUC of 1, while a classifier with random selection has an AUC of 0.5. In reality, a typical classifier has AUC between 0.5 and 1. In selecting operation points on ROC, often times the costs associated with misclassification of each type must be considered when trying to balance the sensitivity with the specificity. This balance is usually adjusted through class priors or class weights. For example, to use the proposed method as a screening tool, we may want to trade (or sacrifice) some specificity for higher sensitivity. This is because if a truly disordered speaker were misclassified as a normal speaker, he or she may miss the opportunity to be further examined by a speech specialist.

2.2.10 Classification techniques

In this work, four different classification algorithms, namely LDA, SVM, extreme gradient boosting (XGBoost), and random forest (RF). For LDA, we consider the effects of shrinkage, a form of regularization to avoid overfitting, along with class priors, to address the unequal costs of misclassification. For SVM, we examine the effects of different kernels, along with class weights. Throughout the modeling procedure, grid search and random search are used for hyperparameter tuning using the *scikit-learn* library in Python [34]. However, class weights or priors are not tuned automatically, and certain discrete values are considered for the study. We consider several categories based on features extracted using feature engineering and different levels of FS. We compare raw data with different feature categories.

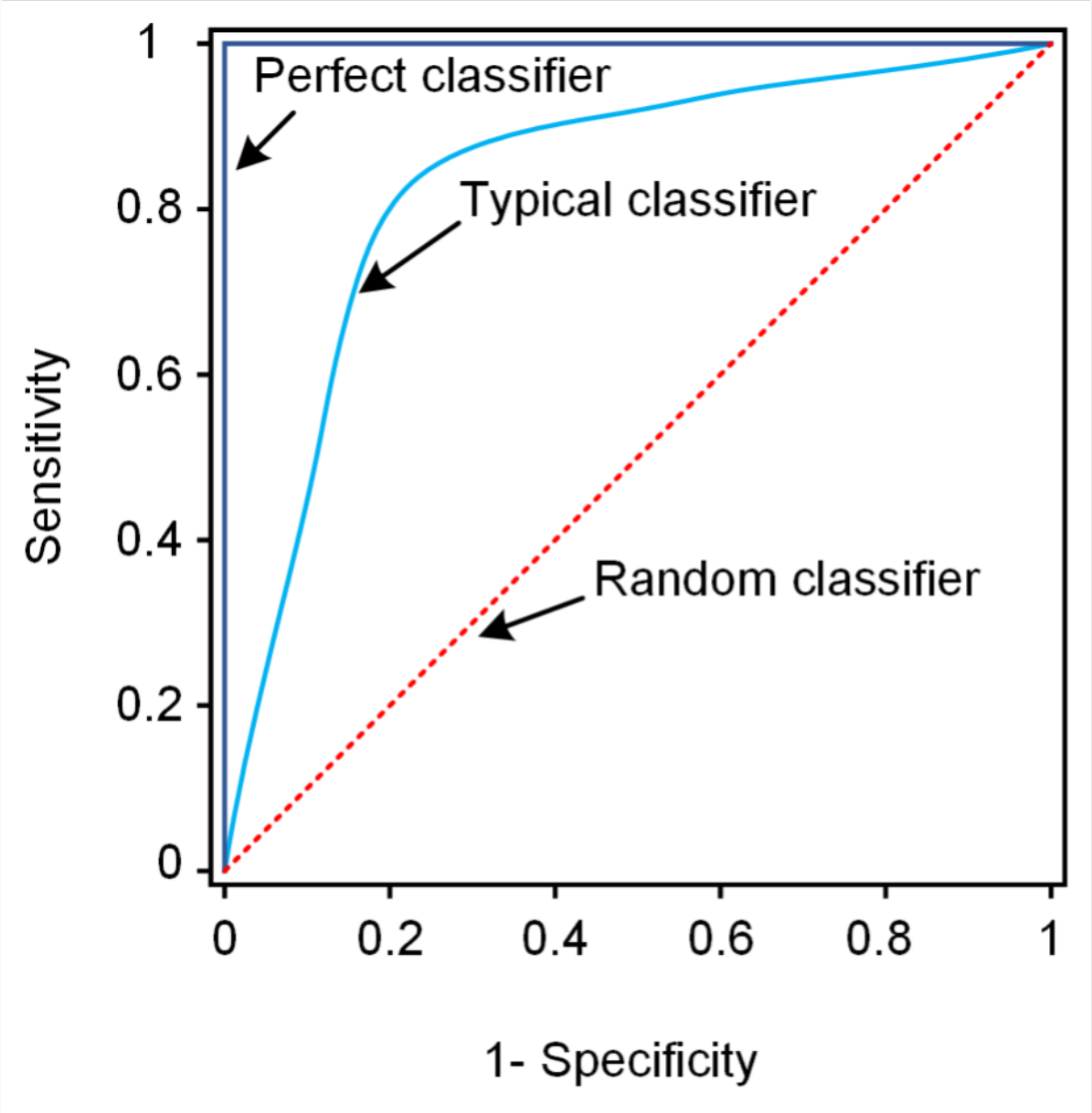

Figure 2.4: The trade-off between sensitivity and specificity illustrated by a ROC curve.

2.2.10.1 Linear discriminant analysis

LDA is one of the most commonly used linear classification techniques used in ML. In this work, the LDA function from Python *scikit-learn* library [35] is used, which generates the linear decision boundary based on the Bayes' rule by modeling the class conditional or posterior probability $P(y = k|x_i)$ of each training sample of d features (i.e., $x_i \in R^d$) for each class k:

$$P(y = k|x) = \frac{P(x|y = k)P(y = k)}{P(x)} = \frac{P(x|y = k)P(y = k)}{\sum_l P(x|y = l)P(y = l)} \tag{2.5}$$

where y is the class label and the class k that maximizes the posterior probability is selected.

Shrinkage is a form of regularization used in LDA to improve the estimation of covariance matrices in situations where the number of training samples is small compared to the number of features. The effect of shrinkage is studied in this work. In addition, the prior probability, $P(y = k)$, is studied on its effectiveness in addressing the unequal costs of misclassification. More information can be found in [34].

2.2.10.2 Support vector machine

SVM is a classification approach developed in the 1990s. Various SVMs have shown superior performance in a variety of settings and are often considered one of the best "out of the box" classifiers [36]. For simple interpretation and to reduce the risk of overfitting, in this work we focus on two-class linear SVM. Consider n samples each with d features (i.e., $x_i \in R^d$, $i = 1, ..., n$) and their labels $y_i \in +1, 1$, linear SVM identifies a hyperplane, which is a liner function in the feature space, i.e., $f(x) = \langle w, x \rangle + b$, where w is the coefficient vector, b is a real constant, and $\langle ., . \rangle$ denotes the dot production in the feature space. The hyperplane is placed such that a maximum distance between the two class samples (i.e., the class margin) is achieved. This is equivalent to the following minimization problem [37]:

$$\min_{w,b} \frac{1}{2} w^T w \tag{2.6}$$

$$s.t. y_i(\langle w, x_i \rangle + b) - 1 \geq 0, \forall i \tag{2.7}$$

For a non-separable case, a soft margin is introduced so that the minimization problem becomes

$$\min_{w,b,\zeta} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \zeta_i \tag{2.8}$$

$$s.t. y_i(\langle w, x_i \rangle + b) \geq 1 - \zeta_i, \zeta_i \geq 0, i = 1, .., n \tag{2.9}$$

where C and ζ_i are constants and ζ_i are called slack variables.

More information on SVM and its training can be found here [36, 37]. In this work, *scikit-learn* [35] with LIBSVM [38] library is used to implement linear SVM. For comparison, we also implemented SVM with nonlinear kernels, including polynomial, radial basis function (RBF), and sigmoid kernels.

2.2.10.3 Random forest

RF is an ensemble method of decision tree (DT) algorithms. It is an extension of bootstrap aggregation or bagging of DTs. In bagging, each classifier's training set is generated by random sampling, with or without replacement from all the samples available for training. Individual predictions of each classifier are aggregated based on a hard or soft voting scheme to form a final prediction. However, unlike bagging, RF also involves selecting a subset of input features at each split point in the construction of trees. *Scikit-learn* is used for implementation of RF. Hyperparameters, including number of trees, max tree depth, and number of features considered for the split, are tuned using random search hyper parameter optimization procedure. More information on RF can be found in [36, 39].

2.2.10.4 Extreme gradient boosting

Another DT-based ensemble method used in this work is boosting. In comparison to bagging, boosting approaches combine various homogeneous weak learners and learn patterns sequentially in an adaptive way. Each of the sequential model depends on the previous ones. XGBoost is one of the most popular boosting approaches, which has been used widely and has achieved state-of-the-art results on many ML challenges [40]. XGBoost is an optimized distributed gradient boosting library, which is implemented under the gradient boosting framework. More information on XGBoost can be found in [40].

2.2.11 Effect of adding triage words

The framework described herein utilizes a singular triage word, "Flower" chosen for its superior performance among the 11 triage words proposed by Anderson and Cohen [29]. While this approach demonstrates efficacy, expanding the analysis to include additional words could

enhance detection rates and provide a broader basis for testing children. Given the short length of individual words and the relatively brief duration of recordings, many feature values are represented as zeros. To address this, three distinct analyses are conducted to integrate more words into the framework.

Initially, we compute feature values for all words and aggregate them on a per-individual basis, thereby maintaining consistent feature counts. Subsequently, we leverage the 10 features previously identified for the word "Flower" and compute their values for other words. Finally, all words are merged and treated as a single sentence, enabling the computation of feature values for each individual within this context. Visual representations of word LMs are provided in Table 2.4, while Table 2.5 illustrates the words arranged with adjusted time, simulating their utterance within a single recording.

Following the generation of feature spaces through these three methods, the same framework is applied to each feature space. Using RFECV, we identify the top 10 features and train a LDA classifier to assess performance.

2.3 Results and discussion

In this work, we conduct investigation from two perspectives: (1) comparing classification performance when different feature sets are used, and (2) comparing classification performance when different classification techniques are used. When comparing different features, the following three feature sets are studied: (a) the original 21 features directly obtained from the SpeechMark Toolbox, which include the counts and strengths of the ten LMs (listed in Table 2.1, considering both onset and offset) for each sample, plus one syllabic count per sample; (b) the 189 features based on rational feature engineering with different feature types listed in Table 2.2 and after redundant features (i.e., Pearson correlation coefficient greater than or equal to 0.99) removed; and (c) The ten features selected from the 189 features via RFECV as discussed in Section 2.2.5. These ten features are listed in Table 2.3. When comparing different classification techniques, they are applied to all the three feature sets.

It is worth noting that all the results presented in this work are based on the unseen test data (i.e., they are not involved in any training steps such as FS or hyperparameter tuning). Due to

Table 2.4: LMs detected for a person uttering 11 triage words

Time	LM	Strength	Time	LM	Strength	Time	LM	Strength
Word 1			**Word 2**			**Word 3**		
0.230	+b	1.0000	0.215	+b	0.7139	0.039	+b	1.0000
0.292	+g	1.0000	0.276	-b	-0.6153	0.108	+b	0.6565
0.302	-v	-0.7102	0.373	-f	-0.6964	0.148	+g	1.0000
0.559	+v	-0.4664	0.380	+g	1.0000	0.459	-s	-0.5000
0.900	-g	-1.0000	0.964	-g	-1.0000	0.460	-g	-1.0000
0.908	+f	-0.1697	1.082	-b	-0.5709	0.612	+g	1.0000
						0.643	-v	-0.5489
						0.852	-g	-1.0000
Word 4			**Word 5**			**Word 6**		
0.366	+b	1.0000	0.086	+b	0.5841	0.060	+b	1.0000
0.404	+g	1.0000	0.092	+g	1.0000	0.132	+g	1.0000
0.786	+v	0.7184	0.216	+s	0.6743	0.420	-g	-1.0000
0.796	-g	-1.0000	0.524	-s	-0.6924	0.435	-b	-0.8939
			0.556	-g	-1.0000	0.474	+b	0.8025
			0.639	+b	1.0000			
Word 7			**Word 8**			**Word 9**		
0.132	+g	1.0000	0.063	+b	0.8336	0.069	+b	1.0000
0.137	+s	0.4957	0.216	-f	-0.6416	0.178	-b	-0.5684
0.212	+s	1.0000	0.220	+g	1.0000	0.204	+g	1.0000
0.420	-g	-1.0000	0.346	-s	-0.6300	0.404	-g	-1.0000
0.502	+b	0.5116	0.396	-g	-1.0000	0.460	+g	1.0000
0.516	+g	1.0000	0.470	+b	1.0000	0.596	-g	-1.0000
0.620	-g	-1.0000	0.532	+g	1.0000			
0.677	-b	-0.8904	0.564	-g	-1.0000			
0.711	+b	0.7589						
0.747	-b	-0.7593						
0.787	+b	0.8804						
Word 10			**Word 11**					
0.111	+b	1.0000	0.05	+b	0.9713			
0.132	+g	1.0000	0.20	+g	1.0000			
0.420	-g	-1.0000	0.45	-s	-0.8637			
0.525	-b	-0.6760	0.50	+s	1.0000			
0.591	+b	0.6158	0.94	-g	-1.0000			

the small number of samples, we would not have enough data for model training if $20 \sim 30\%$ of the dataset were left out for testing as usually recommended. As a result, we leave one sample out from each class for the test set (i.e., totally 2 samples in the testing set). To avoid bias or cherry-picking due to the small number of the test samples, we perform 50 MCVT and use the average of the 50 MCVT runs for performance evaluation. As we have demonstrated

Table 2.5: LMs detected for a person uttering 11 triage words put together in a sentence format

Time	LM	Strength
0.230	+b	1.0000
0.292	+g	1.0000
0.302	-v	-0.7102
0.559	+v	-0.4664
0.900	-g	-1.0000
0.908	+f	-0.1697
1.123	+b	0.7139
1.184	-b	-0.6153
1.281	-f	-0.6964
1.288	+g	1.0000
...	...	...
6.809	+b	1.0000
6.830	+g	1.0000
7.118	-g	-1.0000
7.223	-b	-0.6760
7.289	+b	0.6158
7.339	+b	0.9713
7.493	+g	1.0000
7.742	-s	-0.8637
7.785	+s	1.0000
8.229	-g	-1.0000

previously, this method provides robust and fair evaluations even with small number of samples [33].

As shown in Table 2.6 and Figure 2.5, when the 21 raw features are used, SVM with RBF kernel provides the best overall classification performance with 75.0% accuracy (i.e., 75.0% of the samples are classified correctly). SVM with linear kernel provides the second-best result with 71.0% accuracy. The overall performances of all methods, linear or nonlinear, are relatively poor, indicating that the raw features are not very informative in classifying the two classes.

Table 2.6: Comparison of classification performance based on raw features

Method	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**
LDA	64	54	59
SVM (Linear)	68	74	71
SVM (Poly)	78	32	55
SVM (RBF)	70	80	75
SVM (Sigmoid)	80	54	67
XGBoost	50	86	68
RF	28	76	52

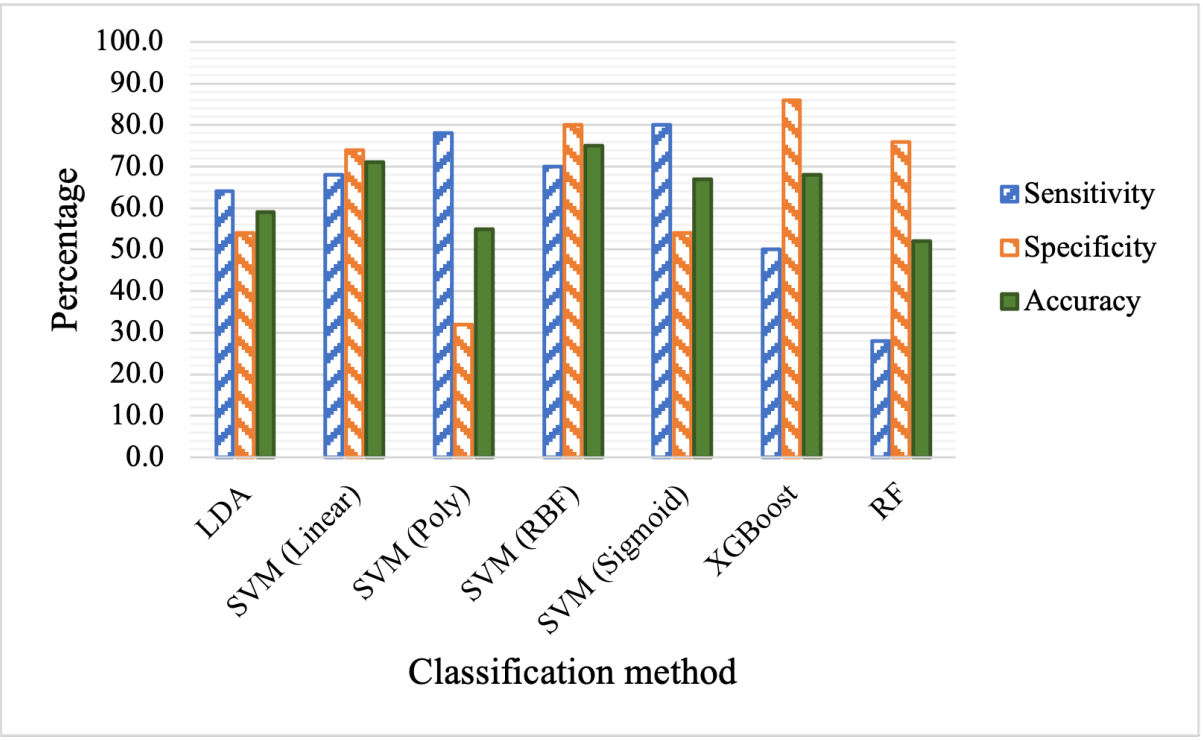

Figure 2.5: Comparison of classification performance based on raw features.

Next, we apply different classification methods to the selected ten features obtained through rational feature engineering and selection. The results are listed in Table 2.7 and shown in Figure 2.6. By comparing Table 2.6 and Table 2.7 (or Figure 2.5 and Figure 2.6), we can see that the performances of all methods have improved in terms of sensitivity, specificity and accuracy. While some improvements are moderate, such as those of SVM (RBF) and XGBoost (with less than 10% improvement in accuracy), others are significant (with as high as 34% improvement in accuracy). Recall that out of the ten selected features, nine of them are newly proposed features. The notable improved performance with these features across all classification methods demonstrates that the proposed features are more informative than the raw features. Since there are seven features that are ratio based, the improved performance is most likely due to our hypothesis that ratio-based features are better at addressing individual variations of samples from the same class. The direct comparison of accuracy using the two sets of features are shown in Figure 2.7. In particular, LDA classifier achieves 94.0%, 92.0% and 93.0% in sensitivity, specificity and overall accuracy respectively. Several other methods also achieve nearly 90.0%

in sensitivity, specificity and overall accuracy, including SVM with linear, polynomial and sigmoid kernels. In addition, using raw features has led to skewed or imbalanced sensitivity and specificity in several methods as shown in Figure 2.5. For example, SVM with polynomial and sigmoid kernels have high sensitivity but poor specificity, while XGBoost and RF have high specificity but poor sensitivity. In comparison, the sensitivity and specificity based on the selected engineered features are much more balanced.

Table 2.7: Comparison of classification performance based on rationally engineered and selected features

Method	Sensitivity (%)	Specificity (%)	Accuracy (%)
LDA	94	92	93
SVM (Linear)	86	92	89
SVM (Poly)	84	94	89
SVM (RBF)	72	94	83
SVM (Sigmoid)	88	88	88
XGBoost	60	88	74
RF	76	94	85

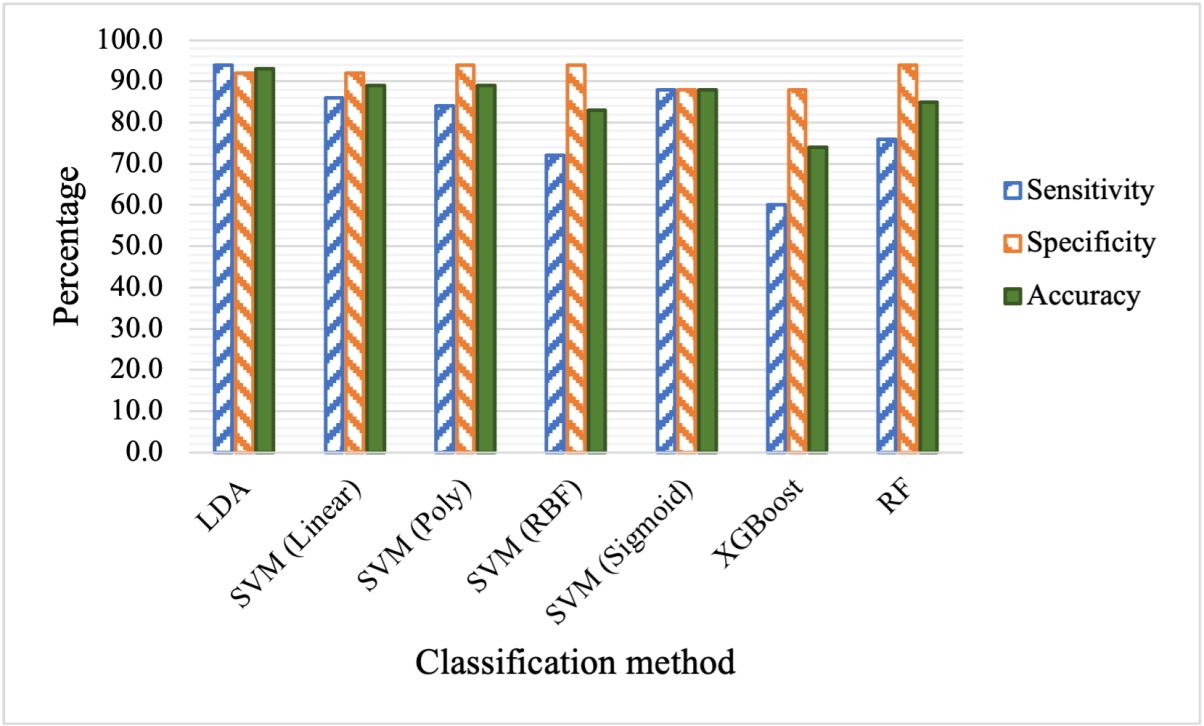

Figure 2.6: Comparison of classification performance when selected features are used.

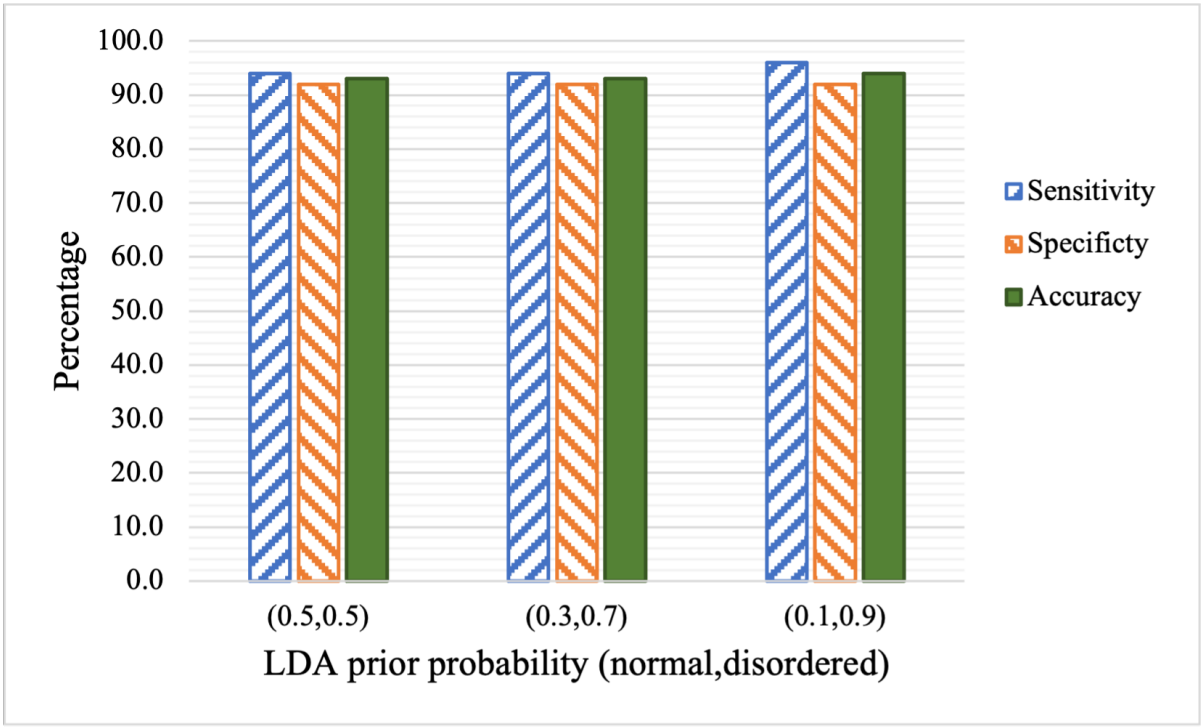

Figure 2.7: The impact of class priors on sensitivity, specificity and accuracy of the LDA classifier.

As discussed previously, class prior probability or class weight can have an impact on sensitivity and specificity, which can also be adjusted to count for the unequal costs of misclassification (i.e., the cost of false positive vs. that of false negative). In this work, since we aim to develop a screening method, high sensitivity is more desirable as the false negative (i.e., children with speech disorder are misclassified as normal speakers) may miss the opportunity to be examined by a speech specialist. On the other hand, false positives will cause less harm other than the cost associated with the follow up examination. Since LDA performs the best among all methods and it is more robust than some of the other nonlinear methods, we focus on examining the impact of class priors on LDA. Three different class priors are studied, namely (0.5,05), (0.3, 07), and (0.1, 0.9). (0.5,0.5) indicates equal priors for normal and disordered classes. (0.1, 0.9) indicates eight times higher prior probability for the disordered class than the normal control group. The expectation is that the priors of (0.1, 0.9) would lead to higher sensitivity compared to (0.5, 0.5) or (0.3, 0.7). The results are shown in Table 2.8 and Figure 2.7, which indicate that the sensitivity and specificity are not significantly affected by the class priors. Specifically, there is no change in sensitivity and specificity when class priors are changed

from (0.5,0.5) to (0.3, 0.7). There is slight increase in sensitivity (from 94.0% to 96.0%) when class priors of (0.1, 0.9) are used, while specificity is unchanged.

Table 2.8: LDA classification performance when different class priors are used

Prior	**Sensitivity**	**Specificity**	**Accuracy**
(0.5,0.5)	94.0	92.0	93.0
(0.3,0.7)	94.0	92.0	93.0
(0.1,0.9)	96.0	92.0	94.0

More thorough examination of the trade-off between sensitivity and specificity the LDA classifier is shown in the ROC curve (Figure 2.8), which is obtained by changing class priors in a much wider range than the three cases presented previously. As discussed previously, the best possible classification method would yield a point in the upper left corner of the ROC space, representing 100% sensitivity and 100% specificity. As shown in Figure 2.8, with proper tuning, the ROC curve of LDA approaches that point with 96% sensitivity and 92% specificity. The ROC curve can serve as a visual tool for selecting LDA tuning parameters, i.e., class priors, based on cost/benefit analysis of the speech disorder screening decision making.

When combining the features from all 11 words per individual, the total number of features remains constant, however with a less proportion of zero features. Following the removal of highly correlated features, a set of 373 features persists. Utilizing Recursive Feature Elimination (RFE) in conjunction with LDA employing SVD solver, 10 features are selected with class priors set at (0.5,0.5). The resulting sensitivity and specificity of the features for the LDA classifier are 100% and 70%, respectively. Notably, altering the class priors does not yield performance enhancements.

In the alternative approach, we focus on extracting feature values specifically for the "Flower" word. However, due to the elimination of highly correlated features in each word's analysis, not all of the originally selected 10 features are retained. Out of an anticipated 110 features, only 36 persist, with the remainder being removed due to correlation with other features. Despite employing various metrics in model training, the optimal sensitivity and specificity achieved for an LDA (LSQR) classifier with class priors set at (0.5,0.5) are 80% and 46%, respectively. Subsequently, 10 features are selected using RFE with LDA (LSQR) under identical

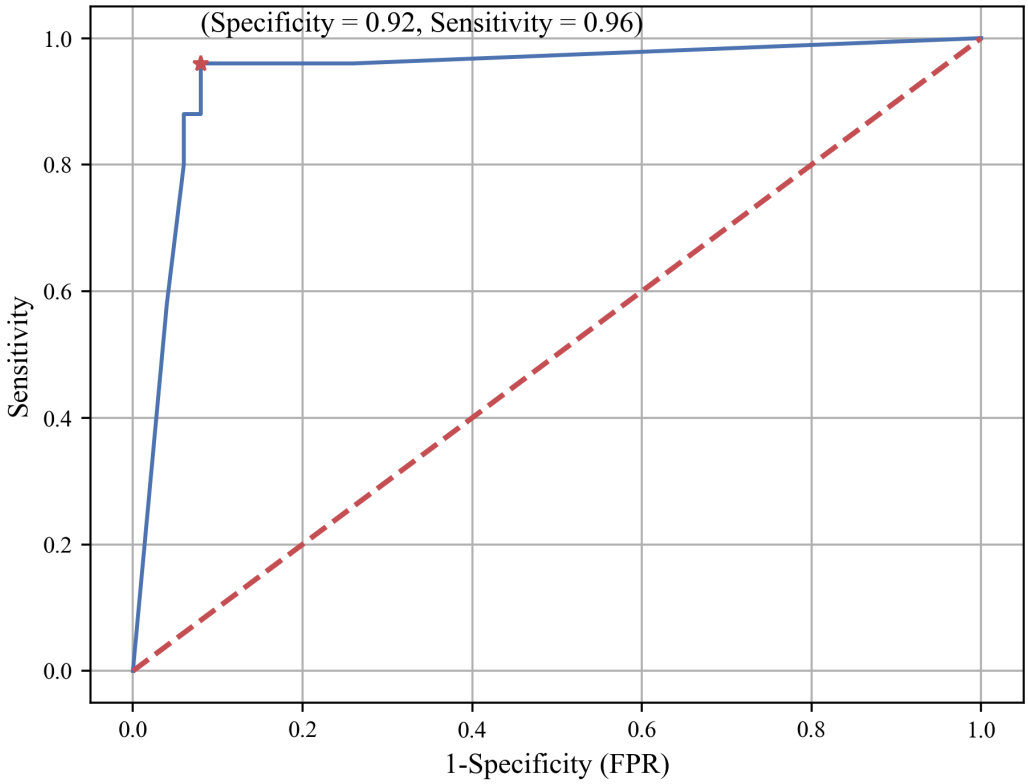

Figure 2.8: ROC curve of the LDA classifier.

class priors, resulting in a notably poor performance with sensitivity and specificity values of 84% and 42%, respectively.

In the final step, wherein all words are combined and treated as a single sentence, the feature set is further refined by eliminating highly correlated features. Among the resulting 1784 features, 1068 redundant features are removed. Of the 716 remaining features, 10 are selected via RFE with LDA (SVD) using class priors of (0.5,0.5). Consequently, the sensitivity and specificity elevate to 96% and 82%, respectively.

The results indicate that integrating more words reduces the model's specificity. Nonetheless, considering the study's prioritization of sensitivity over specificity, the inclusion of extra words is recommended as it may improve detection rates. Broadening the scope to include more words enhances the likelihood of detecting speech disorders, as there is a possibility that a child may exhibit disordered speech in ways not captured by the initial selection of words. By

considering a wider range of words, there is a greater chance of capturing instances of speech disorders that may have otherwise gone undetected. Through hyperparameter tuning, sensitivity levels of up to 100% can be achieved. However, this improvement comes at the expense of specificity, which drops to 56% with the third method. This trade-off underscores the importance of careful consideration by clinicians and specialists, who can determine the appropriate balance between maximizing detection rates and minimizing false positives.

2.4 Conclusion

In this work, we propose an automated computer-assisted screening method for children with speech disorders. The main contribution of this work is to propose a set of novel knowledge-based features that have not been proposed before and to demonstrate the effectiveness of these new features in detecting childhood disorder. In particular, this work proposes specific and average strength of n-grams and ratio-based features. The ratio-based features have been found particularly informative in characterizing audio recordings for speech disorder detection. Similar to the idea of BMI metric used in obesity studies, the ratio-based features proposed in this work are hypothesized to better address the usually wide individual variations among samples from the same class than their individual components. This is validated by the results that significant improvements in classification are obtained based on these new features across different classification methods when compared with those based on the raw features. Similar to many other medical studies where the sample size of normal controls is significantly greater than that of patients, creating so-called sample imbalance problem that can negatively affect many conventional classification techniques. In this work, we found that SMOTE is an effective and easy to implement technique to address this issue. However, cautions must be taken to ensure that the synthesized samples mimic the true minority samples in terms of feature properties (e.g., n-gram counts can only take non-negative integers). In addition, to avoid overfitting, the synthetic samples should not be used as test samples. To improve classification performance and reduce the risk of overfitting, as well as to reduce model complexity, in this work we propose a two-step FS procedure. In the first step, the highly correlated and redundant features are removed through the evaluation of their Pearson correlation coefficients. In the second step,

RFECV is utilized to further reduce the number of features. Through this two-step procedure, it is found that only ten features are needed to obtain the optimal CV accuracy. Based on the raw features and the selected ten features, a systematic study and comparison of different linear and nonlinear ML classification techniques is conducted. It is found that with raw features, all classification methods, linear or non-linear, fail to achieve high classification performance. In comparison, with the ten selected features, which contain nine features proposed in this work, the performances of all classification methods are significantly improved, indicating that the proposed features are more effective for characterizing speech disorder using speech LMs.

It is worth noting that small sample size has always been a limitation in biomedical studies due to labor, time, and other constraints. However, with careful separation of model training (including FS and hyperparameter tuning) and testing (using samples completely left out of the model training process), significant conclusions can be drawn from analyses based on small number of samples. In this regard, MCVT is a robust technique for comparing different modeling techniques and assessing their performances with small number of test samples. FS is also an effective way to avoid overfitting and reduce test variance, especially when there are more features than observations as in this study. Finally, SMOTE can help alleviate the problem by generating synthetic samples. A word of caution is that the above procedures ought to be limited to the training process on the training samples only to avoid overfitting by FS using test samples and bias introduced by the artificial samples.

Ultimately, we suggest that longer audio recordings may increase the likelihood of successful detection. Expanding the sample size and incorporating lengthier recordings could offer notable advantages to this study. Additionally, we propose a detailed examination of why the word “flower” demonstrates efficacy by specialist and suggest the inclusion of additional words similar to ”flower” for further investigation.

Chapter 3

A Systems Engineering Computer-assisted Biomarker Detection Framework for Autism Spectrum Disorder using Proteomic Data

Autism spectrum disorder (ASD) affects approximately 1 in 44 children in the United States. While behavioral criteria are used as the standard for ASD diagnosis, recent proteomic analyses show metabolic differences in the plasma/serum of individuals with ASD. However, identifying reliable biomarkers for ASD has been challenging due to significant variations in protein levels caused by confounding factors such as age, gender, diet, and comorbidities. To address this issue, we propose systematically generating physically meaningful novel features more resilient to these confounding factors than the original measurements, such as protein levels. We then propose an automated computer-assisted biomarker detection framework that integrates these novel features with a hybrid FS technique, and a linear ML model. The effectiveness of the framework was demonstrated using a dataset of serum samples from 76 TD boys and 78 boys with ASD, aged 18 months to 8 years. Our proposed framework identifies a panel of 8 novel features defined in this work. Using the dataset mentioned above, the proposed method detects ASD with high accuracy -achieving an AUC of 0.95, outperforming the previous study of 0.86 with 9 proteins. In addition to the proteins used as features in previous studies, a novel set of engineered features that includes the ratio of proteins is proposed, which reduces within-class variations due to their resilience to confounding factors. Additionally, our FS technique combines a sequential filter and wrapper method to address their respective limitations. A linear ML model is then developed using training samples and independently tested using a set of hold-out samples. The linear ML model is chosen for its robustness to overfitting and superior interpretability. Our methodology introduces systems engineering principles and techniques to ASD detection research. Specifically, biomarkers beyond the traditional physical trait are

defined to include bio-information that can only be extracted by considering their interactions and correlations. The systems engineering perspective provides additional insights into the ASD mechanism, which can lead to additional discoveries in the future.

3.1 Introduction

ASD is a complex neurodevelopmental condition characterized by deficits in social communication and repetitive behaviors [41]. Its prevalence, currently estimated at 1 in 44 children in the United States [42], poses significant personal, familial, and societal challenges [43]. Given these factors, there is a significant focus on advancing our understanding of the underlying pathobiology of ASD, prevention, early detection and developing practical treatment plans [44]. Diagnostic methods primarily rely on behavioral criteria such as difficulties in communication and social interaction, which can be subjective and challenging to apply to younger children [45]. Early diagnosis is crucial due to the potential benefits of early intervention [46, 47]. Therefore, the identification of biological markers to understand the ASD's underlying causes and associated comorbidities, predict ASD risk, aid in early diagnosis, and guide targeted treatments is highly valuable. Biomarkers can be detected in various bodily sources such as tissues, serum, blood, and urine meaning that collecting body fluids for proteomics analysis is a minimally invasive and cost-effective approach [48]. Such markers would enhance diagnostic accuracy, facilitate early intervention, and potentially lead to improved outcomes for individuals with ASD [43].

Researchers have explored biochemical tests to predict ASD diagnosis. This approach reduces subjectivity, enabling earlier detection and intervention implementation using behavioral techniques which have been proven to help the patients [49]. Employing a biochemical approach in diagnosing ASD carries clinical significance and provides insights into the underlying mechanisms of the condition's etiology [50].

Recent advancements in high-throughput technology have revolutionized biomedical sciences by enabling the exploration of multiple biological layers of a disease using various molecular platforms. Genomics, transcriptomics, proteomics, and metabolomics provide comprehensive insights into the molecular processes underlying normal and diseased cells. Genomics

focuses on the entire genome, while transcriptomics examines gene expression patterns using technologies like RNA sequencing. Proteomics studies the structure, function, and flow of proteins, with mass spectrometry (MS) being a powerful tool for analyzing protein expression. Metabolomics examines small molecules resulting from cellular metabolism [51].

Proteomics, enabled by advancements in protein analysis technology and bioinformatic tools, allows for the quantification and characterization of proteins using advanced laboratory techniques including MS and bioinformatics. This field holds promise in identifying disease-associated protein markers for diagnostic, prognostic, and therapeutic purposes [52].

Two major approaches are employed in proteomics research. The bottom-up approach focuses on molecular biology and aims to identify and characterize a particular biomarker at the molecular level. By understanding the relationship between the biomarker structure/function and its function in the disease development, diagnostic, preventive, and treatment strategies can be developed. Individual biomarkers have yielded disappointing results, highlighting the demand for multiple diagnostic/prognostic markers to enhance test sensitivity and specificity [52].

The top-down approach, on the other hand, adopts a bioinformatics perspective. It involves generating proteomic spectra of biomarkers using MS techniques. These spectra provide information on proteins and the analysis of a large number of proteins from normal and patients creates profiles of mass spectra. To unravel the complex protein patterns in mass spectra, higher-order analysis techniques and data mining are employed. This approach offers advantages by not requiring individual protein purification, identification, and antibody development, facilitating clinical assay development. The top-down proteomic approach coupled with MS and data mining algorithms shows promise for early disease detection [52].

Omics analysis approaches can be broadly categorized into statistical analysis and ML analysis. Statistical analysis, such as correlation analysis, provides a descriptive overview of data distribution, aiming to identify patterns and differentially behaving variables within biomedical datasets [53]. On the other hand, ML methods train computational models to extract information, recognize patterns, and make informed decisions. ML techniques are particularly valuable in handling complex data where traditional statistical methods may fall short

[54]. Unsupervised learning methods uncover patterns without predefined class labels, while supervised learning methods utilize labeled training data for classification and prediction tasks [51].

ASD studies often face limitations in sample size [55]. Small sample sizes are susceptible to outliers, making it challenging to characterize data distributions. This limitation has hindered statistical analyses, resulting in reduced power and accuracy, leading to false positives and the omission of important information. In contrast, ML approaches offer promise by applying algorithms to handle small sample sizes and complex datasets without focusing on the knowledge of the dataset and outperforming classical statistical methods [56]. To overcome this limitation MCVT is used to ensures the reliability of classification models. It is worth mentioning that ML methods should complement statistical approaches, not replace them completely.

The task of identifying decisive biomarkers for ASD has proven challenging due to notable fluctuations in protein levels influenced by factors like age, gender, diet, and other comorbidities [57, 58]. To overcome this challenge, we suggest the systematic creation of meaningful novel features that possess physical significance and are more resistant to the impact of these confounding factors compared to the original protein level measurements. The novel set of engineered features includes protein ratios. The use of ratios has been shown to improve classification performance [59, 60] and address the sample variations of individuals in the same class better than using individual protein measurements alone.

Proteomics data faces challenges such as curse of dimensionality, where the number of variables is much higher than the number of samples, leading to computational costs and difficulties in demonstrating statistical significance. The dataset becomes sparser as the number of features increases leading to overfitting. To address these issues, data reduction methods are employed, such as FS, which involves choosing a subset of relevant and non-redundant features. Three main categories of FS methods are filters, wrappers, and hybrid methods, each with its own approach to selecting important features. Filters independently rank features based on a score or correlation, wrappers use ML algorithms iteratively, and hybrid methods combine the first two approaches. These methods offer solutions to the problems of dimensionality and sparsity in proteomics data analysis [61].

In this work, a novel framework is proposed that is consisted of utilizing a new FS criterion to pre-select a subset of features, and incorporating a hybrid FS to further choose a panel of biomarkers. The proposed method is able to select a subset of features with better classification performance of than the previous works on the same dataset. Moreover, the hybrid FS includes a level of randomness that helps with overfitting.

The remaining of this work is arranged into 5 sections. Section 3.2 summarizes the previous efforts in the ASD biomarker discovery techniques, Section 3.3 provides the dataset and the algorithm, Section 3.4 presents the results of the suggested algorithm, and finally, Section 3.5 provides a discussion on the selected features, while Section 3.6 concludes our work.

3.2 Previous works

This section provides a review of previous studies on ASD diagnosis using ML techniques.

In one study, various methods such as Naïve Bayes (NB), K-Nearest Neighbors (KNN), SVM, Logistic Regression (LR), and DL were employed to diagnose ASD in children. These methods were applied to non-clinical datasets specifically designed for ASD screening in children, but not based on blood tests. DL showed the highest accuracy (96-99%) among the tested methods [62]. However, the absence of preprocessing steps like FS and the lack of blood test datasets, limit the findings' generalizability and no discriminating biomarker is discovered.

Many more studies utilized statistical classification methods, including LR, Fisher Discriminant Analysis (FDA), classification and regression trees (CART), and Principal Component Analysis (PCA), for ASD diagnosis, but they do not use blood tests [63, 64, 65, 66, 67, 68]. While these studies highlight the potential of ML for ASD diagnosis, further research is needed to validate the findings in clinical settings with blood test datasets.

Several distinct metabolic differences have been observed in individuals diagnosed with ASD, prompting investigations into their potential roles in the clinical pathology of this condition [50]. Blood and plasma-based metabolite measurements have frequently been utilized for identifying biomarker panels with the potential to predict ASD diagnosis.

In one study, Hewitson et. al analyzed serum samples from 76 boys with ASD and 78 TD boys. Using proteomic analysis, the researchers identified a panel of 9 proteins that showed significant differences between the ASD and TD groups. The panel achieved an AUC of 0.8599, suggesting its potential as a blood biomarker for ASD [43]. Nonetheless, there is still potential for enhancement, and the integration of a pre-selection stage could prove beneficial by eliminating redundant and irrelevant features from the analysis.

A new autism spectrum disorder discovery (ASDD) strategy is developed by Saleh et. al. The strategy includes a novel Hybrid Rejection Technique (HRT) to filter the data. In this work, an Ensemble Technique (ET) with NB, KNN, and DL classifiers are utilized. Experimental results using a dataset of blood tests from children show that the ASDD strategy with an accuracy 0.92 [45]. While this study achieved good performance, it is important to note that outlier rejection should be applied exclusively to the training samples rather than the entire dataset. Additionally, it is advisable to exclude the labels from the outlier rejection technique, as in real-world scenarios, samples will be unlabeled when using the trained model for prediction. Therefore, an algorithm that demonstrates robustness to outliers is required.

In another study, Qureshi et. al used statistical and ML techniques to analyze metabolomic and nutrient measurements in children with ASD compared to TD children. They identified 46 significant nutritional/metabolic differences between the groups and found interconnected metabolic differences in ASD with AUC scores of 0.6–0.9. By using multivariate analysis, they discovered potential biomarker panels consisting of up to six metabolites that could accurately distinguish between ASD and TD with up to 98% predictive accuracy [50]. The drawback of this work is that in their pre-selection stage, they use hypothesis testing which typically assumes a simple relationship between a feature and the target variable, such as a linear relationship. However, in many real-world scenarios, the relationships may be nonlinear or involve complex interactions among multiple features. Hypothesis testing may not adequately capture such complex relationships, leading to the exclusion of important features. Besides, hypothesis testing results can be sensitive to the sample size and the distributional assumptions of the data while Fisher's criterion is relatively robust to changes in sample size since it considers the ratio

of between-class variance to within-class variance. This property makes it applicable even when dealing with small datasets.

3.3 Materials and Methods

3.3.1 Ethical considerations

The research protocol and consecutive modifications were proposed by The Johnson Center for Child Health and Development (Austin, TX), and the study was authorized by the Austin Multi-Institutional Review Board or the IntegReview Institutional Review Board, depending on the date of sample collection. Prior to participation, all participants or their legal guardians provided written informed consent. Participants with any genetic, metabolic, or coexisting physical, mental, or neurological disorder were excluded from the study [43].

3.3.2 Participants

The study enrolled 154 male pediatric individuals, including 76 with a diagnosis of ASD and 78 TD individuals. The mean age of the ASD group was 5.6 years, and the mean age of the TD group was 5.7 years.

For the ASD group, clinical psychologists with research-reliability training utilized the autism diagnostic observation schedule (ADOS) and the autism diagnostic interview–revised (ADI-R) tools to evaluate the participants. Based on the data and DSM-5 criteria, a clinical diagnosis was established. The ADOS diagnostic algorithms were used to measure the overall severity of ASD symptoms.

For the TD group, the adaptive behavior assessment system-second edition (ABAS-II) was used to screen for developmental concerns, and participants with first- or second-degree relatives diagnosed with ASD were excluded from the study.

3.3.3 Dataset

The participants in the study were healthy with no clinical symptoms, and a fasting blood draw was taken from them. The blood was then centrifuged and stored at -80°C. The SomaLogic's

SOMAScan™ $1.3K$ platform was used for examination, which measured 1,317 proteins in $150\,\mu L$ serum in 154 samples [43].

3.3.4 Monte-Carlo cross validation and testing

Like many biomedical datasets, this dataset faces numerous challenges, including the curse of dimensionality— too many variables with a limited number of samples. Despite the richness of variables, not all contribute meaningfully, necessitating the incorporation of domain knowledge to engineer insightful features. However, this expansion of the feature space increases the risk of overfitting. To address this, we employ MCVT process [33]. After feature engineering and preprocessing steps, initially, 20% of samples are randomly selected as test samples and are isolated, and the remaining dataset is further divided into training and validation subsets multiple times for the purpose of modeling training, FS, and hyperparameter tuning. In each MCVT, we pre-select a subset of features based on Fisher's criterion, train LDA classification model, utilize a 5-fold stratified CV approach to identify the optimal hyperparameters for each model, and employ a hybrid FS algorithm. The trained models are then applied to the withheld test samples, where we measure and report their AUC. The MCVT process is repeated 20 times to ensure the reliability of our classification models and evaluate their performance. These measures provide insights into the overall performance of the classifier. A flowchart of each MCVT is shown in Figure 3.1. The feature engineering, preprocessing, and FS are further explained in the next sections.

3.3.5 Feature engineering

This study proposes a novel set of engineered features, which includes the ratios of the proteins in addition to the detected proteins used as features in prior research. This novel feature set is designed to decrease within-class variations due to their resilience to confounding factors. The concept of incorporating ratios is similar to the concept of BMI, which recognizes that weight alone is insufficient to determine if someone is overweight. BMI metric takes into consideration the person's height as well as their weight. Similarly, ratios are often more informative features than individual values alone when it comes to accounting for variations among samples within the same class as shown in [59]. Suthar et. al generated ratios of commonly used features

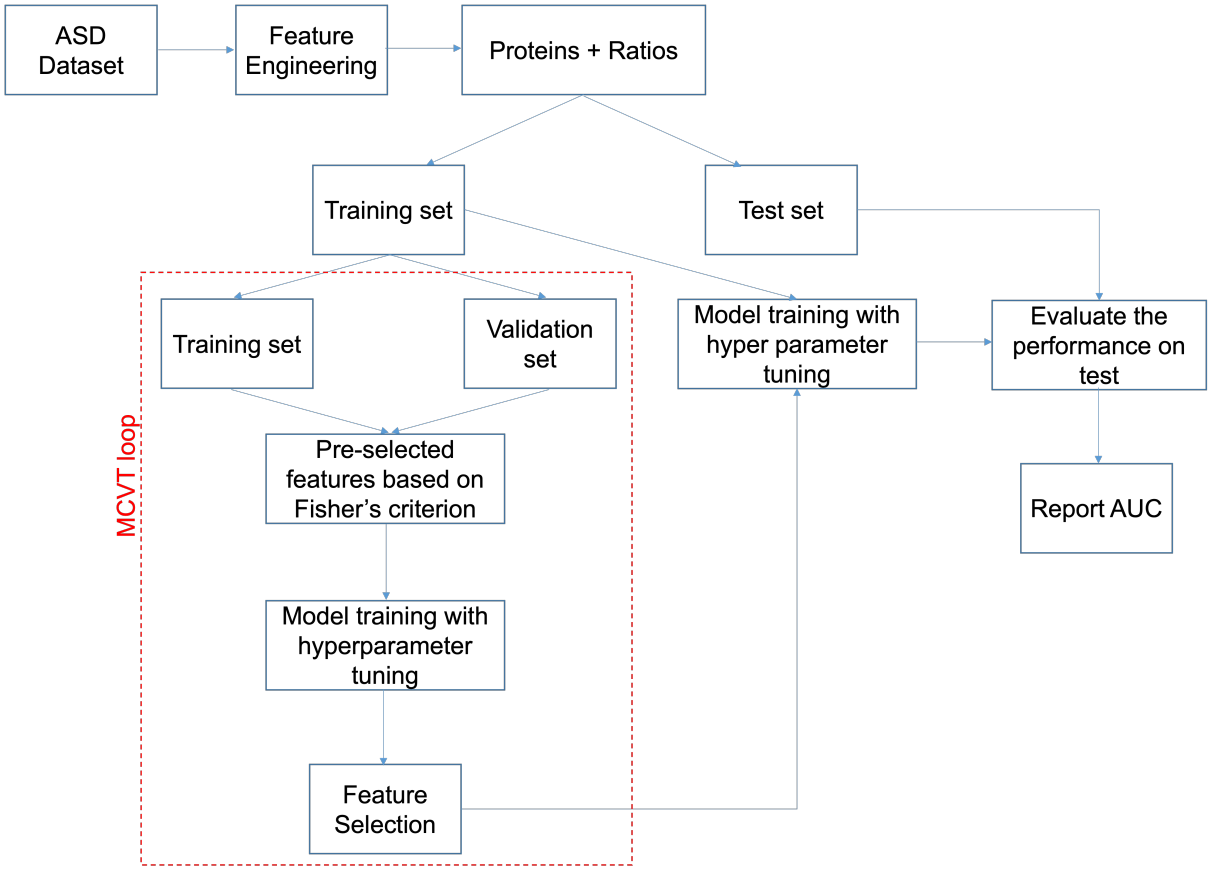

Figure 3.1: Schematic of the MCVT procedure.

in speech deficit studies. Their final panel of 10 features consisted of 7 ratios and resulted in an accuracy of 94%, sensitivity of 96%, and specificity of 92% which suggest that ratio-based features can help in speech disorder detection. This point is further validated by the results that significant improvements in classification are obtained based on these new features (ratios of proteins) when compared to those based on the raw features (proteins) and supports the notion that ratios offer valuable insights.

3.3.6 Data preprocessing

Log2 transformation is commonly applied in bioinformatics and proteomics. Proteomic data often are skewed. Log transformation stabilizes the variance, making the data more homoscedastic and helps in normalizing the distribution of protein expression levels. Additionally, Log transformation rescales data, particularly useful when dealing with wide value ranges, compressing them to enhance visualization and analysis. As a result, the protein levels and protein

ratios were normalized by taking log2 and then z-transformation. The non-normalized proteins were used to generate the ratios. To handle outliers, z-transformed values below -3 and above 3 were adjusted to -3 and 3, respectively. In this work, other transformation methods such as Box-Cox and Winsorizing have been tested which are more robust to outliers.

3.3.7 Feature selection

Numerous studies have demonstrated that the performance of classification techniques can be greatly enhanced by selecting only relevant features as predictors. This is specifically crucial when dealing with a relatively small number of samples compared to the number of features, as it helps to reduce the risk of overfitting and simplify result interpretation. Therefore, FS has become a critical consideration in data-driven approaches. Over the years, a variety of FS techniques have been developed for different modeling and classification applications.

The field of pattern recognition and ML has produced many methods for FS, which can be categorized into three types: filter, wrapper, and hybrid. Filter techniques utilize independent tests to evaluate features, while wrapper techniques need a predetermined learning algorithm. Filter and wrapper methods have their respective advantages and disadvantages. To take advantage of the strengths of both methods and provide a more robust and effective approach to FS, hybrid FS methods have been suggested. In the next 3 chapters we discuss these 3 types of FS techniques.

3.3.7.1 Filter feature selection technique

Filter FS is a popular technique for selecting relevant features from high-dimensional datasets. According to [69], filter methods evaluate features independently of any learning algorithm and select those with the most discriminative power. This approach is computationally efficient and can handle large datasets with ease. The selected features are ranked based on their relevance to the target variable using a statistical or heuristic measure such as correlation, mutual information, or entropy. Filter methods are often criticized for ignoring the interaction between features and the learning algorithm. However, they can complement other FS methods, such as wrapper and hybrid methods, to improve classification performance. Overall, filter FS is a

powerful tool for reducing the dimensionality of high-dimensional datasets and improving the efficiency and effectiveness of ML algorithms.

The filter algorithm begins by selecting an initial feature subset X_0, which can either be an empty set or a randomly selected subset. The algorithm then evaluates the current feature subset X^* with an independent test method (M) and compares it with the best feature subset in the previous step, X_{k-1}. If the new feature subset outperforms the previous one, it is chosen as the current best subset. The algorithm continues until a pre-defined criterion δ is met, which could be one or more of the following: (1) further addition or deletion of any feature does not lead to a better feature subset; (2) the performance requirement is met; or (3) a given limit such as maximum number of iterations or the minimum number of features is fulfilled. Finally, the algorithm outputs the last current best subset, which is X_{best}, that was obtained during the search process.

3.3.7.2 Wrapper feature selection technique

Wrapper FS is a method of selecting a subset of relevant features from a larger set of features in a dataset. In contrast to filter approaches, the wrapper approach evaluates subsets of features using a ML algorithm, such as DT or SVM (A), to measure their impact on the classification performance. Wrapper FS uses the classification performance to guide the search for the optimal feature subset. The method involves an iterative search procedure, starting with an initial feature subset X^*. The classification performance of X^* is then compared with the classification performance of the best feature subset found in the previous iterations. If X^* outperforms the previous best feature subset, then $X_k = X^*$ and evaluating the classification performance of each subset until a stopping criterion δ is reached which has been described in the filter approach previously [70]. The wrapper approach has been shown to be more effective than filter approaches, which use statistical measures to select features, but are not able to capture the interactions between features. However, wrapper methods are more computationally expensive than filter methods and may suffer from overfitting if the dataset is too small. Wrapper FS is a powerful tool for improving classification performance in ML applications and has been

widely used in biomedical research including bioinformatics, genetics, medical diagnosis and other fields [71].

3.3.7.3 Hybrid feature selection technique

Filter-based techniques are widely used due to their ability to efficiently handle high-dimensional datasets and computational speed, as well as their independence from the learning algorithm. However, this method is unable to consider feature dependencies and the interaction with the classifier that leads to varying classification performance when the selected features are applied to different algorithms. In contrast, wrapper approaches are advantageous as they consider feature dependencies and their collective contribution to model generation, resulting in higher classification performance than filter approaches. Nevertheless, the wrapper approaches are more prone to overfitting and can be computationally intensive, especially when processing a large number of features. Thus, the choice of FS technique should be based on the specific application, dataset size, and desired trade-offs between classification performance and computational cost [72].

Hybrid FS methods combine filter and wrapper techniques to take advantage of the strengths of both methods and provide a more robust and effective approach to FS. These methods involve using a filter-based approach to pre-select a subset of features based on statistical measures, such as correlation, and then using a wrapper-based approach to further refine the feature subset using a ML algorithm. Several studies have demonstrated the effectiveness of hybrid FS in various applications, including biomarker discovery [73, 74]. Hybrid FS methods are a promising tool for improving FS performance in ML applications and warrant further investigation.

The hybrid search strategy initiates from a given subset X_0 and integrates a filter approach with a wrapper approach to identify the optimal subsets at increasing cardinality. The filter approach employs an independent test method (M) and a corresponding criterion δ_1 to select candidate features, and the wrapper approach evaluates the candidate features using a particular learning algorithm (A) and a criterion δ_2. Following the identification of the best subset at cardinality k, the overall classification performance is assessed, and if the performance meets

the criterion δ_3, the FS procedure stops with the current best subset of features returned as the optimal feature subset. The search continues otherwise at cardinality $k + 1$ [72].

In this study, numerous filtering and wrapper FS techniques are tested. In the end, a three-step hybrid FS procedure is proposed. In the first step, a filtering FS is proposed in which a subset of ratios are pre-selected based on the Fisher's criterion to help reduce the feature space. In the second step, another filtering and a wrapper FS methods (sequential forward floating search (SFFS)) are used to select features, and this work combines them into a sequential search approach to enhance FS. In the last step, a final panel of features are selected based on their performance on the validation data.

3.3.8 The feature selection algorithm

3.3.8.1 Feature pre-selection with Fisher's criterion

Fisher's Linear Discriminant Analysis (FDA) is a statistical technique utilized to identify the linear combination of variables that best separates two or more groups or classes. The primary goal of FDA is to find a projection of the data onto a lower-dimensional subspace that maximizes the between-class variance and minimizes the within-class variance. This is fulfilled by calculating the Fisher's criterion shown in (3.1), which is the ratio of the between-class variance to the within-class variance and maximizing it.

$$J(w) = \frac{(\mu_1 - \mu_2)^2}{(s_1^2 + s_2^2)} \tag{3.1}$$

The between-class variance measures the degree of separation between the class means, while the within-class variance measures the variability within each class. By maximizing the Fisher's criterion, LDA finds the optimal projection of the data that maximally separates the classes [75].

Fisher's LDA has several advantages over other classification methods such as LR and DTs. It is a linear method, which means it is computationally efficient and can handle high-dimensional datasets. It also assumes that the data is normally distributed, which is often the case in real-world applications.

LDA has abundant applications in diverse fields, such as finance, marketing, genetics, and computer vision. In genetics, LDA has been used for gene expression analysis and disease diagnosis [76].

The Fisher's criterion is calculated for all ratios and are sorted. The selection starts with the feature with the highest score. Each feature contains two proteins. In order to not pick redundant features (i.e., the features that are highly correlated with an existing features), once a ratio is selected, all the ratios that have either of the consisting proteins of the selected ratio are discarded. Since each protein can be present in only one ratio, the number of ratios with non-shared proteins is equal to the half of the number of proteins (each ratio is consisted of two proteins). This process results in 658 ratios with the highest Fisher's criterion. The mentioned procedure lets us to pick ratio of a/b instead of b/a if Fisher's score is higher for the former ratio. Features with higher score theoretically should minimize the within-class variance (tightening the distributions) while maximizing the projections between the means allowing for better separation. In each MCVT, a subset of ratios are selected in this way and proceed to the next step of the FS algorithm.

The procedure of calculating Fisher's criterion and selecting a subset of ratios is shown in Algorithm 1.

After choosing a subset of ratios, we combine them with the panel of 9 proteins suggested by [43] that classifies the ASD participants from TD individuals with an AUC = 0.8599±0.0640 to compare the significance of the added features.

3.3.8.2 The filtering criterion

Pre-selecting a subset of features before applying wrapper approaches can result in more suitable results. As a result, in the next step, the algorithm once again pre-selects a subset of features using an adjusted filtering criterion based on the [72] discussed below. Peng et. al's study introduces a new approach for selecting features in biomedical data classification to handle the challenges posed by high dimensionality.

The algorithm applies LDA to each feature and stores the AUC of training. A feature with higher AUC has better discriminating power. The feature with highest AUC (X_0) is stored

Algorithm 1 Feature pre-selection based on Fisher's criterion

Require: $X_{\text{control}}, X_{\text{case}}$ samples from two classes

1: Compute the means of the two classes:

$$\bar{x}_{\text{control}} = \frac{1}{n_{\text{control}}} \sum_{i=1}^{n_{\text{control}}} x_{\text{control},i}$$

$$\bar{x}_{\text{case}} = \frac{1}{n_{\text{case}}} \sum_{i=1}^{n_{\text{case}}} x_{\text{case},i}$$

2: Compute the within-class scatter matrix:

$$S_w = \sum_{i=1}^{n_{\text{control}}} (x_{\text{control},i} - \bar{x}_{\text{control}})(x_{\text{control},i} - \bar{x}_{\text{control}})^T + \sum_{i=1}^{n_{\text{case}}} (x_{\text{case},i} - \bar{x}_{\text{case}})(x_{\text{case},i} - \bar{x}_{\text{case}})^T$$

3: Compute the between-class scatter matrix:

$$S_b = (\bar{x}_{\text{control}} - \bar{x}_{\text{case}})(\bar{x}_{\text{control}} - \bar{x}_{\text{case}})^T$$

4: Compute the Fisher's criterion:

$$J = \frac{\text{tr}(S_b)}{\text{tr}(S_w)}$$

return J

5: Find J for every feature p and rank them in a descending order

6: Pick q with the highest J

7: Record the proteins in the nominator and denominator of the ratio q:

$$nom \leftarrow q_{nominator}$$

$$denom \leftarrow q_{denominator}$$

8: Remove all ratios that have either nom or $denom$

9: Select the next feature with the highest J

10: **while** $n_{features} < 658$ **do** Repeat steps 7-9

11: **end while**

in the final subset (X_k) and is removed from the whole set (U). The remaining set is named $Y_k = U - X_k$.

The filtering criterion has 3 components as shown in (3.2):

$$T = w_1 * A + w_2 * \mu + w_3 * c \tag{3.2}$$

where w_1, w_2, and w_3 are balancing factors where $w_1 + w_2 + w3 = 1$. A and μ and c are explained below:

Estimated AUCs from trained LDA models are normalized across all features (3.3).

$$A_i = \frac{AUC_i}{\Sigma_{i=1}^{n} AUC_i} \tag{3.3}$$

where n is the number of considered features. Since a higher AUC means better discriminative power of the feature, a higher A is desired.

In order for a feature to be deemed suitable for selection, it should possess a strong ability to discriminate and also complement the existing features within the feature subset. Peng et. al evaluate the level of complementarity between a feature and a group of features through the following estimation [72]:

$$\mu = \frac{\Sigma_{i=1}^{k}(1 - |p_i|)}{k} \tag{3.4}$$

where $p_i \in [-1, 1]$ is the pearson correlation coefficient between the target feature and a feature in the feature subset (X_k). If the training dataset has m samples, the feature vector is $x = (x_1, ..., x_m)$ and a feature in X_k is $y_i = (y_{i1}, ..., y_{im})$, the pearson correlation between the target feature x and y_i is calculated by (3.5):

$$p_i = \frac{m\Sigma_{j=1}^{m} x_i y_{ij} - \Sigma_{k=1}^{m} x_k \Sigma_{j=1}^{m} y_{ij}}{\sqrt{m\Sigma_{j=1}^{m} x_j^2 - (\Sigma_{j=1}^{m} x_j)^2}\sqrt{m\Sigma_{j=1}^{m} y_{ij}^2 - (\Sigma_{j=1}^{m} y_{ij})^2}} \tag{3.5}$$

$(1 - |p_i|) \in [0, 1]$ demonstrates the independence between the target feature x and y_i. This value for two dependent features is equal to 1. μ measures the (average) complementarity of a target feature to all the features in feature subset X_k. The higher the value of μ, the more complementary feature x is to the features in subset X_k.

The correlation of a feature to the label is crucial in ML because it indicates the relationship between the feature and the target variable. A high correlation suggests that the feature carries valuable information and has a strong influence on predicting the target variable. In other words, a feature that is highly correlated with the label is likely to provide meaningful insights and contribute significantly to the learning algorithm's ability to make accurate predictions.

By considering the correlation of features to the label, ML models can identify which features are most relevant and informative for the task at hand. This knowledge enables the models to focus on the most influential features, leading to improved accuracy and predictive performance. Moreover, feature-label correlation helps in FS and dimensionality reduction, as it allows for the identification and elimination of irrelevant or redundant features that may introduce noise or unnecessary complexity to the learning process.

Therefore, a third term c is added to measure the connection of a continuous feature x to the binary target variable (dichotomous variable). For this purpose, the point biserial correlation coefficient r_{pb} is calculated using (3.6):

$$r_pb = \frac{M_1 - M_0}{s_n}\sqrt{\frac{n_1 n_0}{n^2}} \tag{3.6}$$

where s_n is the standard deviation, M_1 is the mean value on the continuous variable x in group 1, and M_0 is the mean value on the continuous variable x in group 2. n_1 is the size of group 1, n_0 is the size of group 2 and n is the total sample size [77]. $c \in [0, 1]$ is the absolute value of the point biserial correlation coefficient $r_{pb} \in [-1, 1]$.

The T criterion tries to find a feature that has a good discrimination power by itself (first and third components), and the feature is complementary to the chosen features in X_k.

T values are found for each feature. The T values are then normalized so that the summation of them equals to 1. Features are then ranked based on their normalized T values in ascending order. Conventional filter-based approaches often rely on selecting features based on their rank or applying a threshold to exclude irrelevant ones. However, these methods have a notable drawback since they only consider the top-ranked features, which may not necessarily be the best candidates and will cause the model to overfit. As a result, in contrast, this method selects the features with randomness included to add flexibility. This means when a feature subset (Z_k) is going to be selected, the features with higher values of T would most likely be selected, but the features with lower T values would also have a chance of being selected [72].

After the T values are ranked, we apply a min-max normalization and add λ as shown below to ensure that even the minimum of the normalized T values has a chance for selection.

$$nT = \frac{T - min(T)}{max(T) - min(T)} \tag{3.7}$$

$$probsel = \lambda + (1 - \lambda) * nT \tag{3.8}$$

where nT is normalized T after min-max normalization. $probsel$ is the probability of selection, and $\lambda = 10e^{-5}$ is tuning parameter.

After this, a vector of random numbers with the same length as Ts is generated. The feature with the highest difference is selected. This method ensures that features with highest values are selected most of the time, however, features with low values still have a chance of being selected. This eventually prevents the model from overfitting since a feature with a high T value would not necessarily result in a high test performance and including features with low values might benefit the test performance. This process is repeated 10 times and 10 features are pre-selected.

These pre-selected features are added to X_k (for the first iteration, it consists of the feature with the highest AUC) and are input to the wrapper FS.

3.3.8.3 Sequential feature selection

Recent comparative studies have shown that sequential search methods, particularly the SFFS algorithm, produce classifiers with better or comparable classification performance [78]. The SFFS algorithm is the most preferred choice in many applications, especially when classification reliability is the primary concern [72]. The proposed approach in [72] employs the SFFS algorithm as the searching mechanism.

The sequential search approach is utilized to choose the optimal feature subset by adding or removing a single feature or a few features at a time until specific criteria are met. There are three primary types of search strategies: sequential forward selection (SFS), sequential backward selection (SBS), and bidirectional selection. SFS starts with an empty set and successively adds features until a proper feature subset is found. SBS starts with the complete set of initial features and removes relevant features until the desired feature subset is obtained. The

limitations of SFS and SBS approaches are that after a features is discarded in SFS, it would not be re-selected, and after a feature is selected in SBS, it would not be discarded. Over the years, many variations of sequential searching have been developed, with the most successful approach by Pudil et al. The floating search methodology includes the SFFS and the sequential backward floating search (SBFS). The SFFS and SBFS methods were developed to overcome the "nesting effect" drawback of SFS and SBS, respectively [79, 80].

SFFS algorithm starts with an empty feature subset $X_0 = \phi$. The algorithm includes a feature from the remaining set Y_k with the highest performance based on the J_{X_k}. J_{X_k} is the evaluation function and depending on whether the SFFS is filter-based or wrapper-based it could be an independent test or a ML algorithm, respectively. SFFS later finds the least significant feature in the subset and excluded that until the new subset has a better performance. This process is repeated until the stopping criteria is met. The SFFS procedure is shown in Algorithm 2.

Algorithm 2 SFFS

Require: U: initial set of features, f_{crit}: stopping criterion
Ensure: X_k: optimal feature subset

$X_k \leftarrow \emptyset$
$k = 0$
$Y_k = U$ ▷ The set of remaining features
while $f < f_{crit}$ **do**
    for $c \in Y_k$ **do**
        $J_c \leftarrow J(X_K \cup c)$
    end for
    $y = argmax(J_c)$ ▷ The best feature from the not selected subset
    $X_{k+1} = X_k \cup y$ ▷ Add the best feature to the subset
    for $p \in X_k$ **do**
        $J_p \leftarrow J(X_{k+1} - p)$
    end for
    $x = argmax(J_p)$ ▷ The least significant feature in the subset
    while $J(X_{k+1} - x) > J(X_k)$ **do**
        $X_k = X_{k+1} - x \cup y$ ▷ Remove the least significant feature from the subset
        $k = k - 1$
        for $p \in X_k$ **do**
            $J_p \leftarrow J(X_{k+1} - p)$
        end for
        $x = argmax(J_p)$
    end while
end while
return X_k

After the features are pre-selected by filtering criterion, SFFS selects the best combination of the features (X^*) and their AUC is recorded. The process continues if the new subset outperforms X_k. This whole procedure repeats until the difference of AUCs of X_k and X^* is less than 0.001 or the number of iterations reaches to 20. A subset of features are then selected and the next MCVT starts. After the 20 MCVTs are done, the combination occurs and the final panel of biomarkers is selected. This whole algorithm is illustrated in Figure 3.2.

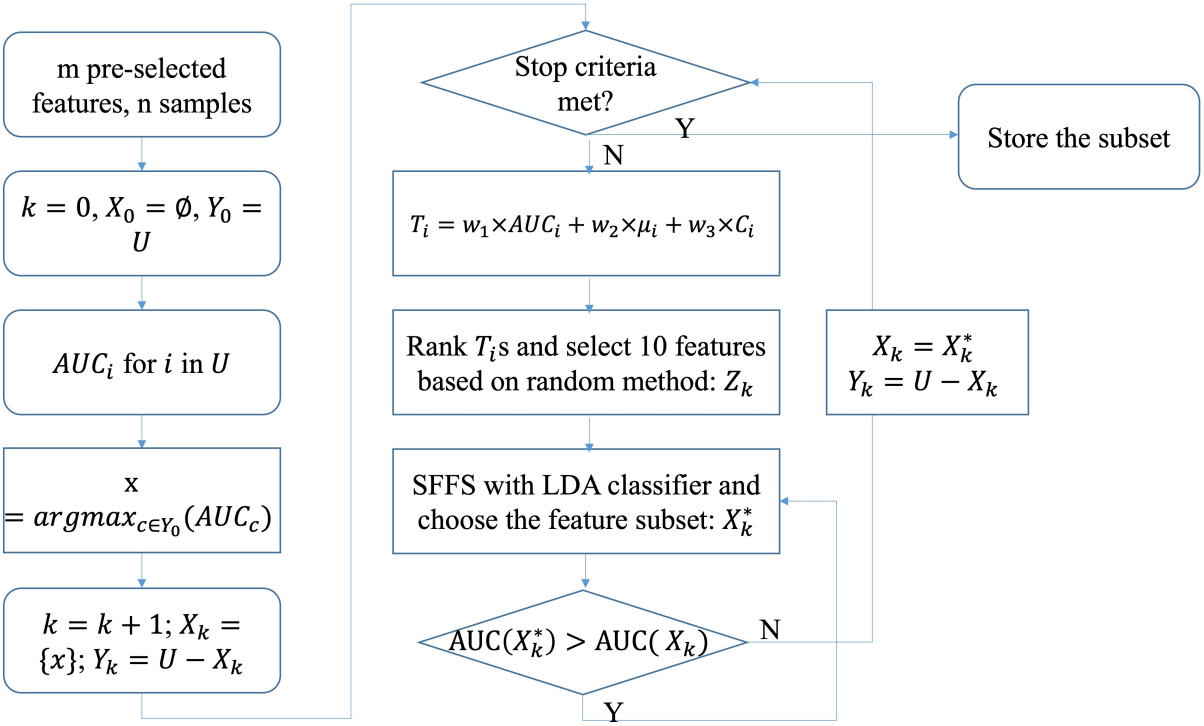

Figure 3.2: The hybrid FS algorithm applied in each MCVT.

3.3.8.4 Final panel of biomarkers

As shown in the flowchart of the algorithm in Figure 3.1, the dataset is split to training and test sets with an 80/20 proportion. The test set is set aside and FS is not based on the test samples in any of the steps. The training set is then split into training and validation sets with the same proportion. This process is repeated 20 times (20 MCVTs). In each MCVT, the hybrid FS technique is applied, a subset of feature is selected, and the AUC of validation is recorded to be used as weight later. Next, the 20 subsets are combined. The features are combined with their

respective weights. The weight for each subset is found by multiplying the AUC of validation of that iteration by 10 and taking the integer, as shown in (3.9):

$$weight_{MCVT} = integer(AUC^{valid}_{MCVT} * 10) \tag{3.9}$$

For instance, if 10 features are selected in iteration 5, and the validation AUC for that iteration using those features is 0.905, the 10 features in this subset are repeated 9 times and then put in the final combination of features. The reason for giving weights to the subsets is that some features might appear in most of the iterations although they result in a low performance. Thus, giving weights to subsets based on their performance ensures that important and significant features will be selected. After all features are combined based on their respective weights, the top 20 most frequent features are selected.

In the final step, to find the final number of features, the features are selected based on their frequency one by one and the performance of those features are found for the 20 MCVTs and stored. The average of AUCs of validation for the 20 MCVTs are plotted against the number of features. To find the final number of features, we employ the one-standard-error rule to choose a model. This involves computing the standard error of the validation AUC for each model size, and then opting for the model where the validation AUC error falls within one standard error of the highest point on the curve. The rationale behind this approach is that if several models appear equally effective, selecting the simplest model- i.e., the one with the fewest predictors would suffice [81].

To report the performance of the final panel of biomarkers, we split the dataset to train (80%) and test (20%) 20 times and train an LDA on the training set with fine-tuning hyperparameters through 5-fold CV and evaluate the performance on the test set. An average of these 20 iterations is reported as the final AUC with respect to that panel of biomarkers. Repeating this process makes the results more trustful and reliable. Given the limited number of samples and the substantial variability within each group (be it cases or controls), the efficacy of the model heavily relies on the selection of training samples. Therefore, if the model demonstrates strong performance using the final biomarker panel, it indicates the discriminative power of

the selected features. To compare the performance of the selected features with the 9 proteins suggested in [43], with the same split, the performance of the 9 proteins is also found.

3.3.9 Classification techniques

In this work, four different classification algorithms, namely LDA, LR, SVM, and gaussian naïve bayes (GNB) is used. For LDA, we consider the effects of shrinkage, a form of regularization to avoid overfitting, along with class priors and different solvers, to address the unequal costs of misclassification. For LR, we explore the effects of l1 and l2 regularization, regularization strength, along with class weights. For SVM, we examine the effects of different kernels, regularization, along with class weights. Finally, for GNB, we consider the effect of adding a smoothing parameter to the variance of all features. Throughout the modeling procedure, grid search and random search are used for hyperparameter tuning using the *scikit-learn* library in Python [34]. However, class weights or priors are not tuned automatically, and certain discrete values are considered for the study. We consider several categories based on features extracted using feature engineering and different levels of FS. We compare proteins suggested by [43] with the selected feature subset.

3.3.9.1 Linear discriminant analysis

To benefit from the advantages of utilizing a hybrid FS method, after ranking features with a filter-based FS method, SFFS is used with a ML model to act as a wrapper method. In this study, LDA is applied as the evaluation function in the SFFS algorithm. At any time, the performance of a feature subset is evaluated, LDA is used.

LDA is a linear classification technique, built on Fisher's linear discriminant, that is widely used in ML. LDA employs the Fisher's criterion to reduce the dimension of the data. LDA maximizes the separation between classes by maximizing the distance between means of two classes and minimizing the within-class variances.

In this study, the LDA function of the Python *scikit-learn* library [82] was utilized. This function generates a linear decision boundary by modeling the posterior class-conditional probability density function (PDF) $P(y = k|x_i)$ for each training sample in each class k, based on Bayes' rule.

$$P(y = k|x_i) = \frac{P(x_i|y = k)P(y = k)}{P(x_i)} \tag{3.10}$$

where x_i is the i-th training sample of d features (i.e., $x_i \in R^d$), y is the class label, and k is the selected class that maximizes the posterior probability. $P(y = k)$ is the prior probability, and $P(x_i)$ is the marginal probability. The class label y is then determined based on the class k that maximizes the posterior probability.

Using Bayes' rule, we can rewrite (3.10) as:

$$P(y = k|x_i) = \frac{P(x_i|y = k)P(y = k)}{\sum_{j=1}^{K} P(x_i|y = j)P(y = j)} \tag{3.11}$$

where K is the total number of classes. The decision boundary is linear, and is given by:

$$y(x) = \text{argmax}_k\{\log P(y = k) + \log P(x|y = k)\} \tag{3.12}$$

where x is a new test sample with d features, and $P(x|y = k)$ is the class-conditional PDF.

In practice, LDA is often used in situations where the number of training samples is small compared to the number of features. In such situations, shrinkage is a form of regularization that can improve the estimation of covariance matrices. Least squares (lsqr) solution solver of the LDA function of the Python *scikit-learn* library can be combined with shrinkage. The effect of shrinkage can be studied by adjusting the shrinkage parameter. In situations where the number of features is larger, singular value decomposition (svd) solver should be used since it does not compute the covariance matrix. In this study, svd solver was used due to the large number of features, and shrinkage is not suitable for svd solver.

3.3.9.2 Logistic regression

LR, a fundamental statistical method widely employed in classification tasks, models the probability that a binary outcome, such as the presence or absence of a condition, occurs given a set of independent variables. This technique, pioneered by David Cox in the 1950s [83], is particularly suited for scenarios where the response variable is binary and the relationship between the predictor variables and the outcome is nonlinear. Mathematically, LR employs the logistic

function, also known as the sigmoid function, to model the probability of the binary outcome. The logistic function is defined as :

$$p(X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p)}}$$

where $p(X)$ represents the probability of the positive outcome, $X_1, X_2, \ldots, X_p$ are the predictor variables, $\beta_0, \beta_1, \ldots, \beta_p$ are the model coefficients, and e is the base of the natural logarithm.

In this work, *LogisticRegression* function form *scikit-learn* library is used with regularization.

3.3.9.3 Support vector machine

SVM is a recognized classification method developed in the 1990s, renowned for its superior performance across diverse domains and often regarded as one of the most effective "out of the box" classifiers [36]. To ensure interpretability and mitigate the risk of overfitting, we narrow our focus in this study to the two-class linear SVM. Given a dataset comprising n samples, each characterized by d features ($x_i \in \mathbb{R}^d$, $i = 1, \ldots, n$), and their respective labels $y_i \in \{+1, -1\}$, the linear SVM seeks to identify a hyperplane, represented as a linear function in the feature space, i.e., $f(x) = \langle w, x\rangle + b$, where w denotes the coefficient vector, b is a constant term, and $\langle \cdot, \cdot \rangle$ denotes the dot product in the feature space. The placement of this hyperplane is optimized to maximize the margin between the samples of the two classes, a task formulated as the following minimization problem [37]:

$$\min_{w,b} \frac{1}{2} w^T w \tag{3.13}$$

$$\text{s.t.} \quad y_i(\langle w, x_i\rangle + b) - 1 \geq 0, \quad \forall i \tag{3.14}$$

In scenarios where the data are not linearly separable, a soft margin approach is introduced, transforming the optimization problem into:

$$\min_{w,b,\zeta} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\zeta_i \tag{3.15}$$

$$\text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \quad i = 1, \ldots, n \tag{3.16}$$

Here, C and ζ_i are constants, with ζ_i referred to as slack variables, allowing for a soft penalty on misclassifications.

Further insights into SVM and its training methodologies are available in prior literature [36, 37]. In our study, we utilize *scikit-learn* [35] coupled with linear and nonlinear kernels, including RBF.

3.3.9.4 Gaussian Naïve Bayes

GNB is a simple yet effective probabilistic classification algorithm based on Bayes' theorem with the "naïve" assumption of feature independence. It is particularly well-suited for datasets with continuous features and is widely used in various fields such as text classification and medical diagnosis [84]. In GNB, each feature is assumed to follow a Gaussian (normal) distribution within each class. Mathematically, the class-conditional probability density function $P(X_i \mid y)$ of feature X_i given class y is modeled as:

$$P(X_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_{iy}^2}} \exp\left(-\frac{(X_i - \mu_{iy})^2}{2\sigma_{iy}^2}\right)$$

where μ_{iy} and σ_{iy}^2 are the mean and variance of feature X_i in class y, respectively. Given a set of features $X = \{X_1, X_2, \ldots, X_d\}$, the probability of a sample belonging to class y is computed using Bayes' theorem as:

$$P(y \mid X) = \frac{P(y)\prod_{i=1}^{d} P(X_i \mid y)}{P(X)}$$

where $P(y)$ is the prior probability of class y, and $P(X)$ is the evidence or marginal likelihood. Despite its simplicity, GNB often performs well in practice, especially when the

independence assumption approximately holds [85], or when we have limited samples, making it a popular choice for baseline classification tasks.

3.3.10 Receiver operator characteristic curve

Sensitivity and specificity are two most commonly used critical metrics when dealing with binary classification problems in healthcare. Sensitivity is the true positive rate, i.e., the classifier's ability to detect diseased patients correctly, and specificity is the true negative rate, i.e., the classifier's ability to detect normal controls (i.e., the ones without diseases) correctly. We also use accuracy as a single measure when we need to evaluate the overall performance of a classifier. The mathematical definitions of these terms are given below.

In healthcare, binary classification problems are common, and sensitivity and specificity are two of the most crucial metrics used. Sensitivity measures how well the classifier correctly identifies patients who have the disease, while specificity measures how well the classifier identifies patients who do not have the disease. Accuracy is another measure that provides an overall evaluation of a classifier's performance. Below are the mathematical definitions of these metrics.

$$\textit{Sensitivity} = \frac{TP}{TP + FN} \tag{3.17}$$

$$\textit{Specificity} = \frac{TN}{TN + FP} \tag{3.18}$$

$$\textit{Accuracy} = \frac{TP + TN}{TN + TP + FN + FP} \tag{3.19}$$

where TP is the number of true positives (correctly classified positive samples), FN is the number of false negatives (incorrectly classified negative samples), TN is the number of true negatives (correctly classified negative samples), and FP is the number of false positives (incorrectly classified positive samples).

In binary classification, there is often a trade-off between the ability of a classifier to detect true positives (sensitivity) and its ability to avoid false positives (specificity). To visualize this trade-off, a ROC curve is used, which is a useful tool to evaluate classifier performance. In

biomedical informatics, ROC curves are frequently used to assess the effectiveness of classifiers [86, 87, 88]. ROC curve plots sensitivity against (1 - specificity) as illustrated in Figure 3.3.

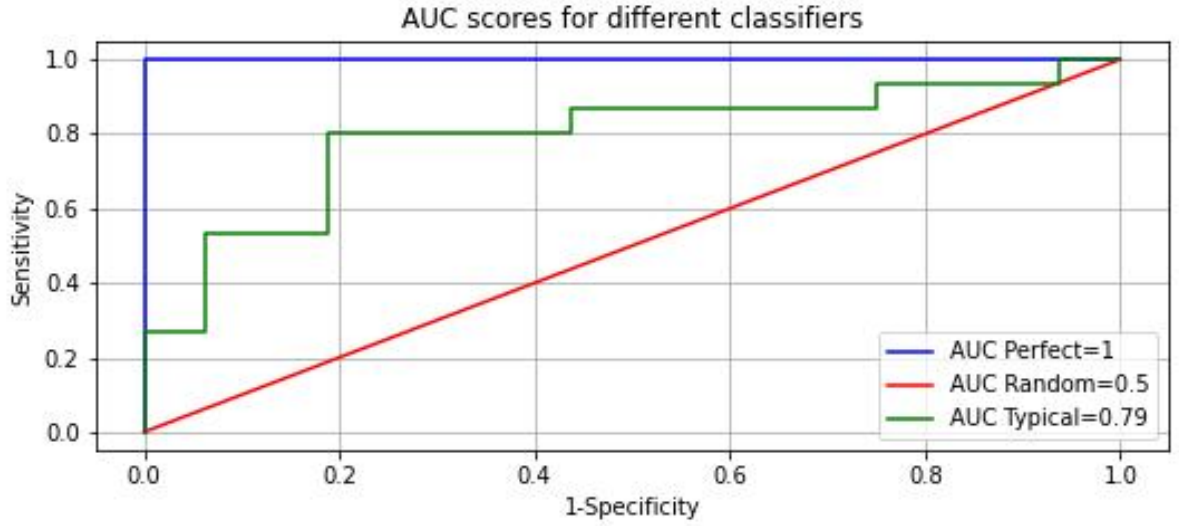

Figure 3.3: The ROC curve illustrating the trade-off between false positive and negative rates.

The AUC provides a single measure of the classification capability of a classifier. A perfect classifier has an AUC of 1, while a classifier with random selection has an AUC of 0.5. Most classifiers have AUC values between 0.5 and 1 as shown in Figure 3.3. A classifier with a higher AUC area performs better than one with a lower AUC area. To find the best trade-off between sensitivity and specificity, we need to consider the costs associated with misclassification of each group. For instance, in disease detection the cost of misclassification a diseased patient as a normal individual should be higher than the cost of misclassification of a normal individual as a diseased one. We can balance the sensitivity and specificity by adjusting class priors or class weights.

AUC considers the overall quality of the associated features and the balance between their sensitivity (true positive rate) and specificity (true negative rate). It takes into account the classifier's ability to distinguish between positive and negative instances across various classification thresholds. By incorporating the complete range of true positive and false positive rates, the AUC provides a robust indication of the feature's discriminative power.

In this research, the AUC is adopted as a more reliable measure to estimate the discriminative capability of each feature. This allows for a more thorough evaluation of the features and enables the identification of those that contribute significantly to the classification task.

3.4 Results

3.4.1 Feature engineering

The novel features generated in this study is the ratios of the 1317 proteins present in the dataset. The number of ratios can be found by $\frac{m*(m-1)}{2}$, where m is the number of proteins. If ratio of a/b is found, the b/a is not considered due to redundancy. 866586 ratios are generated as a result. In this work, we combine the ratios with the panel of 9 proteins found by [43] and the rest of the solo proteins are not considered.

3.4.2 Feature selection

3.4.2.1 Pre-selection based on Fisher's criterion

Since there are a lot of ratios, a pre-selection method is applied using Fisher's criterion. 658 ratios with non-sharing proteins were selected. As a result, the dataset with 667 features (9 selected proteins by [43] and 658 ratios) is fed to the hybrid FS algorithm.

Based on the Figure 3.4, applying the one-standard-error rule using the validation set results in the selection of the 8 variable model. All the selected features are ratios with a test AUC of 0.95. The features in the panel are listed in Table 3.1. To prove the power of ratios instead of raw proteins, the Fisher's score is found for the selected features and reported in Table 3.1. Now, if we take the constituent proteins of all the ratios selected by the algorithm and find the Fisher's score for them, they are smaller compared to the ratios suggesting that ratios have a higher discrimination power. The scores for the constituent proteins are shown in the Table 3.2.

To further prove the potential of ratios, we find the scores for all proteins and all ratios with non-sharing proteins and choose the top 10 for each category. As shown in Table 3.3 and Table 3.4, the top 10 ratios have higher scores compared to the top 10 proteins.

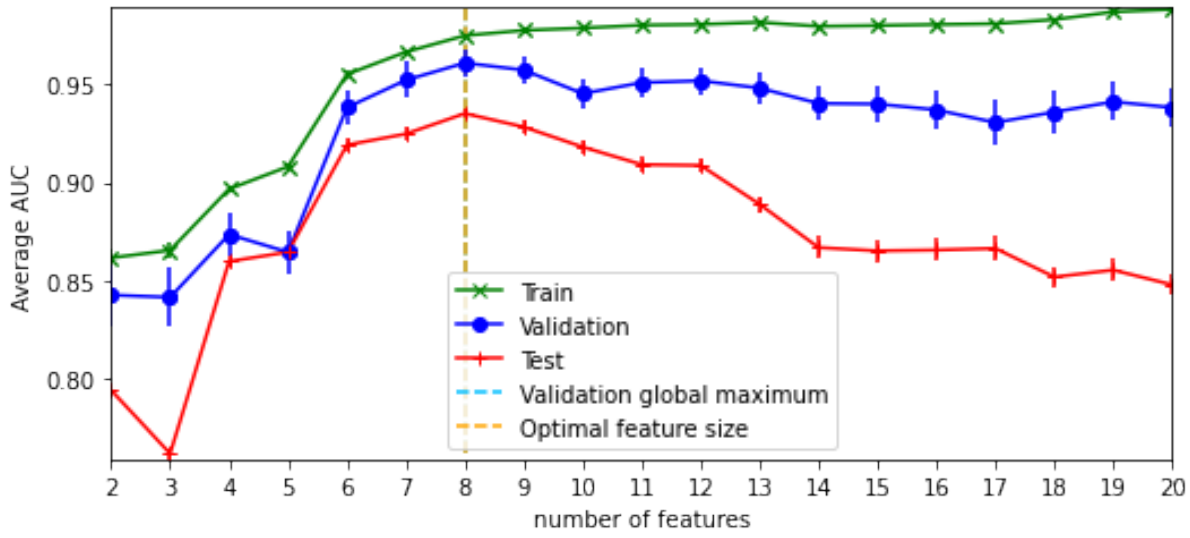

Figure 3.4: The plot of average AUC for a range of number of features. The selection is based on the AUC of validation (red plot). The global maximum of this curve is shown with a green dotted line. The optimal model size is shown with a yellow dotted line.

Table 3.1: Selected features for training data and their Fisher's scores

Ratio	**Fisher's score**
DERM/IL-1Rrp2	0.7406
IgD/FAM3D	0.5415
C5b, 6 Complex/Angiopoietin-2	0.4525
DR6/IL-6 sRa	0.4253
SRCN1/DPP2	0.3922
b2-Microglobulin/M-CSF R	0.3506
a-Synuclein/Aflatoxin B1 aldehyde reductase	0.3312
hnRNP K/P-Cadherin	0.2982

Additionally, the fisher's scores of the 9 proteins suggested by [43] are listed in Table 3.5. The features have much smaller scores compared to the ratios selected by the algorithm.

Finally, to further examine the power of ratios, the feature-label correlation for the panel of 8 features has been found using point biserial correlation. The results illustrated in Table 3.6 and Table 3.7 suggest that the correlation coefficients are slightly higher for the selected panel compared to when we combine all the constituent proteins. The correlation coefficients for the 9 proteins in [43] listed in Table 3.8 suggest that the selected ratios have more correlation to the label.

Table 3.2: Fisher's scores for the constituent proteins of the selected ratios

Protein	**Fisher's score**
DERM	0.4789
IgD	0.3947
Angiopoietin-2	0.1921
DR6	0.1687
P-Cadherin	0.1554
b2-Microglobulin	0.1492
DPP2	0.1325
FAM3D	0.1237
hnRNP K	0.1151
C5b, 6 Complex	0.0918
SRCN1	0.0692
Aflatoxin B1 aldehyde reductase	0.0667
IL-1Rrp2	0.0472
M-CSF R	0.0365
IL-6 sRa	0.0356
a-Synuclein	0.0166

Table 3.3: Fisher's scores for the top 10 ratios with the highest scores

Ratio	**Fisher's score**
DERM/CD59	0.7552
EPHB2/IL-18 Ra	0.6840
C1QR1/Ephrin-A2	0.5464
IgD/FAM3D	0.5415
PTN/IL-1Rrp2	0.5004
ENTP5/MIA	0.4785
IL-17 RC/RELT	0.4587
C5b, 6 Complex/Angiopoietin-2	0.4525
Nectin-like protein 2/Dtk	0.4320
DR6/IL-6 sRa	0.4253

Table 3.4: Fisher's scores for the top 10 proteins with the highest scores

Protein	**Fisher's score**
DERM	0.4789
IgD	0.3947
C1QR1	0.3707
PTN	0.3393
RELT	0.3205
MRC2	0.2870
Calcineurin	0.2844
MIA	0.2822
OMD	0.2809
EPHB2	0.2776

Table 3.5: Fisher's scores for the 9 proteins suggested in [43]

Protein	Fisher's score
DERM	0.4789
IgD	0.3947
EPHB2	0.2776
ROR1	0.2312
MAPK14	0.1971
suPAR	0.1534
eIF-4H	0.0944
ARSB	0.0694
GI24	0.0672

Table 3.6: Point biserial correlation coefficients of the selected ratios with label

Ratio	Coefficient
DERM/IL-1Rrp2	0.5203
IgD/FAM3D	0.4612
C5b, 6 Complex/Angiopoietin-2	0.4288
DR6/IL-6 sRa	0.4183
SRCN1/DPP2	0.4057
b2-Microglobulin/M-CSF R	0.3862
a-Synuclein/Aflatoxin B1 aldehyde reductase	0.3759
hnRNP K/P-Cadherin	0.3586

Table 3.7: Point biserial correlation coefficients of the constituent proteins of the selected ratios with label

Protein	Coefficient
DERM	0.4400
IgD	0.4058
Angiopoietin-2	0.2952
DR6	0.2789
P-Cadherin	0.2690
b2-Microglobulin	0.2635
DPP2	0.2499
FAM3D	0.2419
hnRNP K	0.2316
C5b, 6 Complex	0.2097
SRCN1	0.1833
Aflatoxin B1 aldehyde reductase	0.1797
IL-1Rrp2	0.1521
M-CSF R	0.1339
IL-6 sRa	0.1321
a-Synuclein	0.0908

3.4.2.2 Hybrid feature selection

667 features including 9 proteins and 658 ratios are the fed to the algorithm. The parameters of this algorithms are specified below: 1. The size of the pre-selection subset Z_k is equal to 10

Table 3.8: Point biserial correlation coefficients of the 9 proteins suggested in [43] with label

Protein	**Fisher's score**
DERM	0.4400
IgD	0.4058
EPHB2	0.3489
ROR1	0.3218
MAPK14	0.3003
suPAR	0.2680
eIF-4H	0.2132
ARSB	0.1832
GI24	0.1808

($|Z_k| = 10$). 2. The weights for the filtering criterion are $w_1 = 0.7$, $w_2 = 0.2$, and $w_3 = 0.1$. Different ranges of parameters were tested, and the best parameters were selected based on the validation AUC. 3. In SFFS, 5-fold stratified CV with LDA classifier is utilized to check the performance of the feature subset. *GridSearchCV* Python function from *sklearn* package is applied for finding the optimal parameters. AUC is the metric used for evaluation of the performance. 10-fold CV does not improve the performance by much and results in higher computational cost. The *SequentialFeatureSelector* function is imported from *mlxtend* Python package with floating and forward parameters set to *True* to make is a floating forward FS. 4. Every place LDA is used, the hyperparameter tuning is done with 5-fold CV. The *LinearDiscriminantAnalysis* function from *sklearn* python package is used with SVD solver. Changing class weights does not affect the performance. 5. After feature subsets from all MCVTs are combined with their respective weights, top 20 most frequent features are selected. 6. For different numbers of features from 2 to 20 (from the top 20 most frequent feature subset), the training set is split into training and validation for 20 times. The average of validation AUC of these 20 iterations for each number of feature is plotted against the number of feature. The global maximum of this plot determines the number of features in the panel of biomarkers. 7. To evaluate the performance of the final panel, we find the AUC of train and unseen test. In another way, the original dataset (with 154 samples) is split into training and test samples with an 80:20 ratio, 20 times, to show that the performance of the panel would be high no matter what set of training data is used. An LDA classifier is trained on the training dataset and the AUC of the test samples are stored. Features with a correlation coefficient of higher than 0.9

are eliminated. The mean of all 20 iterations is reported as the AUC of the respected panel of biomarkers. 8. The dataset is split into training and test samples with an 80:20 ratio. The training set is then split to training and validation set with the same ratio. This ensures that the test samples have not been used for features selection and the classification model training.

3.4.3 Performance evaluation

In order to show the power of our algorithm, a LR, LDA, SVM, and GNB classifiers are trained on the dataset with the proposed panel of biomarkers in [43] for 20 MCVTs. The performance of the proposed feature subset is evaluated by the mean and standard deviation of the AUC of the test set in the 20 iterations. In Table 3.9, the results of different classifiers with the 9 proposed proteins in [43] are illustrated. In comparison, the proposed panel of biomarkers using the algorithm suggested in our work is shown in Table 3.10. The results show a 11.4-16.5% improvement in AUC among all methods. The standard deviation of the AUCs are also much smaller compared to the panel selected by [43]. Comparison of Table 3.9 and Table 3.10 for LDA suggests that the performance of the proposed panel in [43] varies from as low as 0.77 to 0.93 based on the training samples, while the performance of our proposed panel is consistently high ranging from 0.83 to 1 suggesting that the performance of the panel does not rely on the training samples. The ROC curve for the panel using LR and LDA classifiers is illustrated in Figure 3.5.

In the end, the boxplot of the selected features are plotted in Figure 3.7 and compared to the boxplot of the panel of the proteins selected by [43] as shown in Figure 3.6. The features selected by this work clearly show a better separation further proving the point that the novel ratio-based features helped the discrimination of the two classes.

To show the effect of adding the correlation of features to the label, we repeat the process without the third term in T ($w_1 = 0.7, w_2 = 0.3, w_3 = 0$). The results suggest that the addition of feature-label correlation improves the result slightly using the same number of features (AUC of LDA classifier = 0.950 vs. 0.908).

To check the sensitivity of the algorithm to parameters (w_1, w_2, w_3), we train models three times with only setting one of the parameters to 1 and the rest to 0. a range of w is tested.

Table 3.9: AUC of LR, LDA, SVM, and GNB classifiers on the panel of 9 biomarkers proposed by [43]

MCVT	LR	LDA	SVM	GNB
1	0.8458	0.8792	0.8250	0.8708
2	0.8875	0.9042	0.8917	0.8708
3	0.7542	0.8208	0.7250	0.6708
4	0.6864	0.7818	0.6273	0.6318
5	0.7583	0.8167	0.7417	0.7375
6	0.9000	0.8708	0.8667	0.9250
7	0.8866	0.8992	0.8992	0.8319
8	0.8504	0.8205	0.8547	0.8419
9	0.7983	0.8193	0.7815	0.7983
10	0.8318	0.8682	0.8227	0.9364
11	0.7137	0.7735	0.6752	0.6752
12	0.8000	0.8125	0.7917	0.8333
13	0.8361	0.8319	0.8571	0.7815
14	0.8419	0.8504	0.8162	0.8803
15	0.9060	0.9017	0.9103	0.8462
16	0.9083	0.9208	0.8917	0.8458
17	0.9076	0.9160	0.8613	0.9034
18	0.9444	0.9316	0.9444	0.8803
19	0.8875	0.8708	0.8750	0.8583
20	0.6708	0.7875	0.6750	0.6500
mean	0.8308	0.8539	0.8167	0.8135
std	0.0791	0.0491	0.0877	0.0926

Results in Table 3.11 suggest that AUC plays a more important role than feature-feature and feature-label correlation.

The initial division of the dataset into training and untouched test sets is replicated 20 times to assess the performance of the proposed algorithm thoroughly. The outcomes presented in this study are derived from the seed showing the most consistent results, and the features from that specific iteration are recommended in the final panel of biomarkers.

3.5 Discussion

The selected features in this work are DERM/IL-1Rrp2, IgD/FAM3D, C5b, 6 Complex/Angiopoietin-2, DR6/IL-6 sRa, SRCN1/DPP2, b2-Microglobulin/M-CSF R, a-Synuclein/Aflatoxin B1 aldehyde reductase, and hnRNP K/P-Cadherin. Some of the constituent proteins of these ratios have shown to be effective in ASD detection.

Table 3.10: AUC of LR, LDA, SVM, and GNB classifiers on the panel of 8 biomarkers proposed by our algorithm

MCVT	LR	LDA	SVM	GNB
1	0.9750	0.9708	0.9833	0.9542
2	0.9375	0.9500	0.9333	0.9333
3	0.9458	0.9542	0.9500	0.9667
4	0.9773	0.9636	0.9773	0.9682
5	0.9292	0.9375	0.9333	0.9250
6	0.9417	0.9500	0.9292	0.9333
7	0.9622	0.9706	0.9496	0.9538
8	0.9402	0.9402	0.9444	0.9444
9	0.9454	0.9370	0.9412	0.9370
10	0.9818	0.9864	0.9864	0.9818
11	0.9402	0.9316	0.9487	0.9444
12	0.9167	0.9042	0.9000	0.8417
13	0.9286	0.9370	0.9286	0.9244
14	0.9487	0.9487	0.9487	0.9573
15	0.9701	0.9701	0.9701	0.9573
16	0.9958	0.9958	0.9958	0.9958
17	1.0000	1.0000	1.0000	1.0000
18	0.9872	0.9615	0.9701	0.9829
19	0.9958	0.9792	0.9917	0.9875
20	0.8417	0.8333	0.8417	0.8167
mean	0.9530	0.9511	0.9512	0.9453
std	0.0363	0.0363	0.0371	0.0459

Table 3.11: AUC of the algorithm with different parameters

Parameters	# Features	Train AUC	Test AUC
$w_1 = 1, w_2 = 0, w_3 = 0$	10	0.9751	0.9417
$w_1 = 0, w_2 = 1, w_3 = 0$	12	0.9169	0.7583
$w_1 = 0, w_2 = 0, w_3 = 1$	10	0.9627	0.9000

Dermatopontin (DERM) protein is primarily found in connective tissues playing a role in the organization and maintenance of the extracellular matrix providing structural support to various tissues in the body. While there is limited research exploring the involvement of DERM in ASD, some studies have indicated that ASD model of mouse shows alterations and loss of the extracellular matrix [89]. Changes in the composition or function of the extracellular matrix could potentially impact the overall brain development, which are areas of interest in understanding the etiology of ASD. Although further research is needed to determine the specific mechanisms by which DERM may be involved in ASD pathogenesis, this protein has been identified as a potential biomarker for ASD detection in previous works [43, 45].

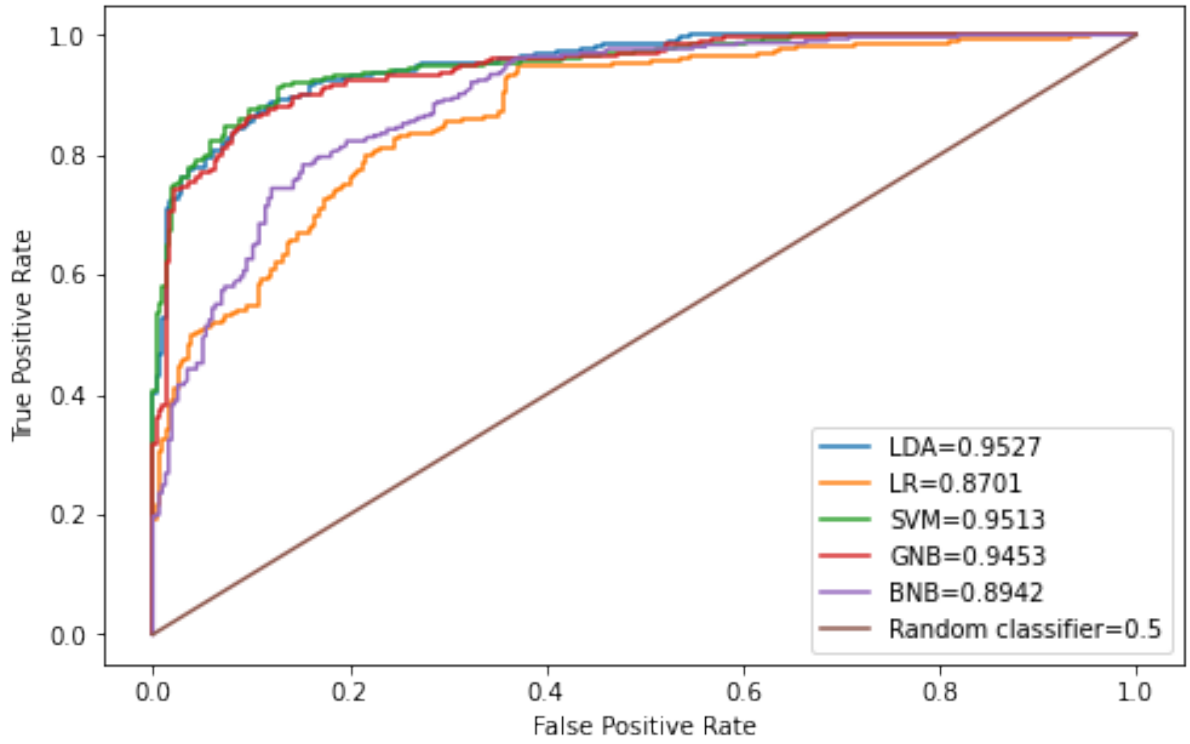

Figure 3.5: The ROC curve of the 8 suggested features using LDA, LR, SVM, and GNB classifiers.

Interleukin-1 receptor-like 2 (IL-1Rrp2) protein acts as a receptor for the interleukin-1 (IL-1) cytokine family involved in immune and inflammatory responses. IL-1Rrp2 has been identified as a potential genetic risk factor. Genetic variations may impact the functioning of the immune system and the regulation of inflammatory processes, potentially contributing to the development and manifestation of ASD. However, no research has demonstrated the precise mechanisms by which IL-1Rrp2 influences ASD.

The precise function of Immunoglobulin D (IgD) protein in relation to ASD is not fully comprehended, and its specific involvement in ASD pathogenesis is not well understood. However, there is emerging evidence suggesting immune response alterations and/or inflammatory pathways in children with ASD. IgD, as part of the immune system, is involved in immune responses and antibody production. Perturbations in immune response pathways have been implicated in ASD, and some studies have reported differences in immune markers, including IgD, between individuals with ASD and TD individuals [43, 45].

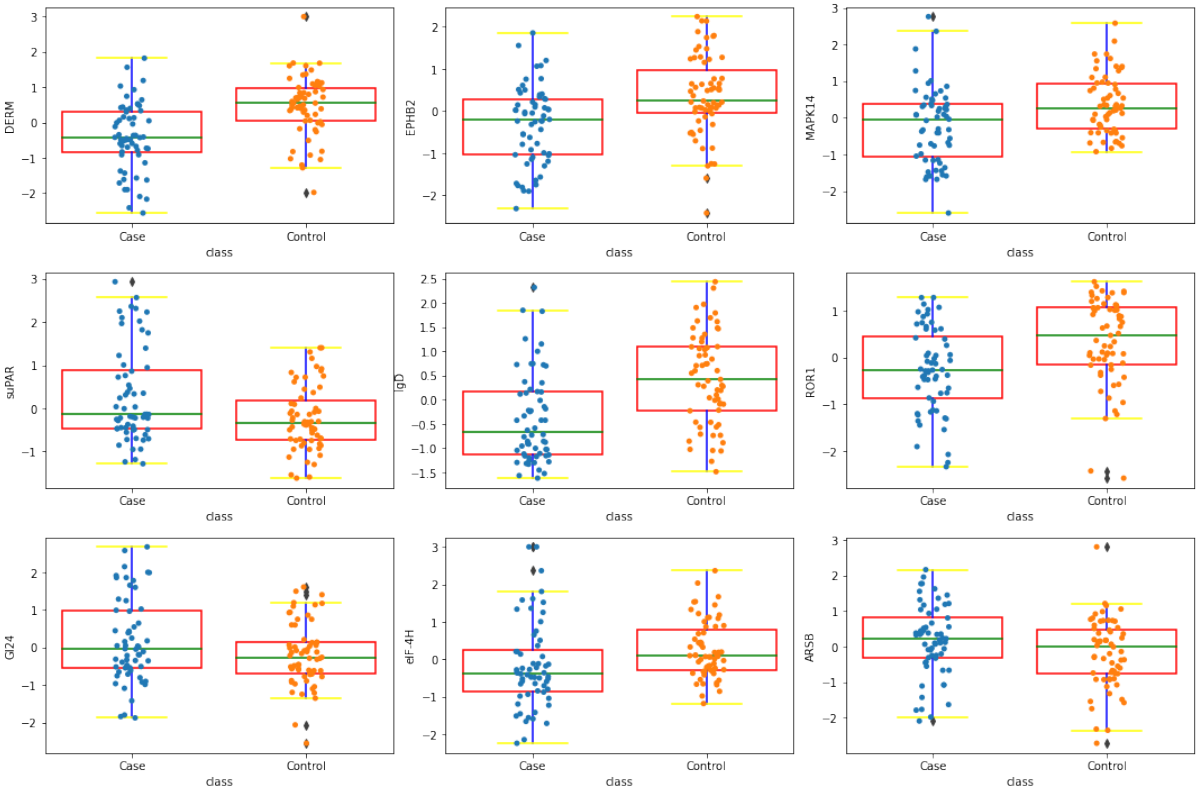

Figure 3.6: The boxplot of the 9 proteins suggested by [43]. The median line is colored in green, whisker lines in purple, and the minimum and maximum lines in yellow. The cases are colored blue while the controls are orange. The outliers are colored in black.

Family with sequence similarity 3 member D (FAM3D) is a protein that belongs to the FAM3 family and plays a role in cell signaling pathways, particularly in metabolic regulation. FAM3D is associated with neurological functions and disorders, particularly within the context of mild ASD. It is also associated with schizophrenia within the mild ASD subgroup [90]

The Complement C5b-C6 complex (C5b, 6 Complex) protein is a crucial component of the complement system involved in immune responses, particularly in the formation of the membrane attack complex (MAC). Upon activation, C5b-C6 associates with other complement proteins to form MAC, which leads to the lysis of target cells by disrupting their membranes. MAC's involvement in neurodevelopmental disorders such as ASD is suggested in [91]

Angiopoietin-2 (Ang-2) is a protein that regulates angiogenesis, the formation of new blood vessels, and contributes to vascular integrity, permeability, and responses to physiological and pathological stimuli in the body. Changes in angiogenesis dynamics and vascular function have been observed in various neurodevelopmental disorders, including ASD [92].

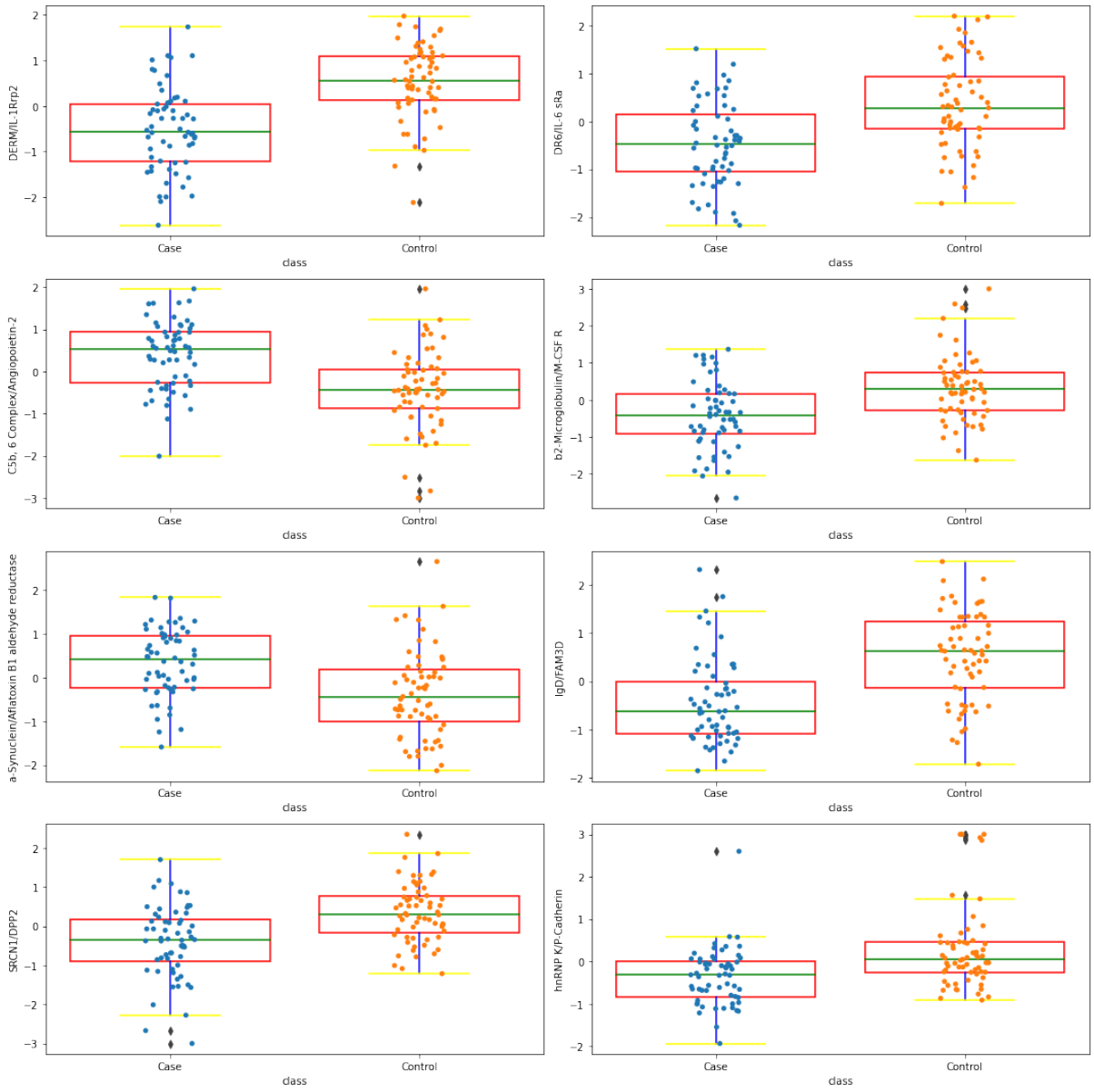

Figure 3.7: The boxplot of the 8 features suggested by our algorithm. The median line is colored in green, whisker lines in purple, and the minimum and maximum lines in yellow. The cases are colored blue while the controls are orange. The outliers are colored in black.

While specific studies linking Ang-2 to ASD is limited, it is plausible that alterations in Ang-2 levels or activity could contribute to vascular abnormalities observed in ASD.

Death Receptor 6 (DR6) protein belongs to the tumor necrosis factor (TNF) receptor superfamily that plays a significant role in various cellular processes, including cell survival, apoptosis, and immune regulation. DR6 can modulate immune responses and neuronal survival resulting in neurodegenerative diseases, depression, and ASD [93].

Interleukin-6 receptor subunit alpha (IL-6 sRa) is a protein involved in immune response regulation and inflammation. A study suggested a potential link between immune response and IL-6 expression in the context of ASD [94]. In another work by [95] finding suggests that elevated levels of pro-inflammatory cytokines, including IL-6sR, may be associated with improved social behavior and reduced impairments in individuals with ASD, nonetheless additional research is needed to fully understand the role of these cytokines in ASD.

SRC kinase signaling inhibitor 1 (SRCN1) protein belongs to the Src family of kinases, which are involved in regulating various cellular processes, including cell growth, differentiation, and communication. In a study, SRCN1 was reported to down-regulate in depression models indicating that SRCN1 may be involved in the pathophysiology of depression [96]. But the specific role of SRCN1 in ASD is not studied before.

Dipeptidyl peptidase 2 (DPP2) protein belongs to the dipeptidyl peptidase family, which plays a role in the breakdown of proteins. Studies have identified alterations in protein metabolism and enzyme activities in individuals with ASD [97, 98]. Therefore, it is plausible that DPP2, along with other proteins involved in protein metabolism, could be relevant in ASD. However, no research suggested the potential relationship between DPP2 and ASD detection.

Beta-2-microglobulin (b2-Microglobulin) protein plays a crucial role in the immune system, specifically in the function of major histocompatibility complex (MHC) class I molecules. These molecules are responsible for presenting antigens to immune cells, contributing to immune surveillance and defense mechanisms. While the direct relationship between b2-Microglobulin and ASD is not extensively studied, alterations in immune system functioning have been implicated in ASD. More research is necessary to understand the role of b2-Microglobulin in ASD pathogenesis and its significance in the disorder.

The role of Macrophage colony-stimulating factor 1 receptor (M-CSF R) protein is primarily regulating the growth, differentiation, and survival of a type of immune cell, macrophages. M-CSF R has been implicated in modulating immune responses and inflammation, which are areas of interest in understanding ASD. Due to altered levels or dysregulation of M-CSF R have been observed in individuals with ASD compared to TD individuals [95], this protein can be a promising biomarker for ASD detection.

Alpha-synuclein (α-Synuclein) protein is primarily found in the brain, where it plays a role in the regulation of neurotransmitter release and synaptic function. It is commonly associated with neurodegenerative disorders such as Parkinson's disease [99]. A study demonstrated variations in α-Synuclein levels in the plasma of children with ASD [100].

Aflatoxin B1 aldehyde reductase member 2 (Aflatoxin B1 aldehyde reductase) protein is an enzyme that plays a role in the metabolism and detoxification of aflatoxin B1, a potent carcinogenic and mutagenic compound produced by certain fungi. Although the direct relationship between Aflatoxin B1 aldehyde reductase and ASD is not well-established in literature, an elevated level of this protein has been reported in neurodegenerative diseases such as Alzheimer's and dementia [101].

Heterogeneous nuclear ribonucleoprotein K (hnRNP K) is a protein that plays a role in various cellular processes, including RNA processing, gene expression, and protein synthesis. While there is limited specific research on the role of hnRNP K in ASD detection, hnRNP K plays a role in the development of the nervous system and signaling of important chemicals in the brain, such as neurotransmitters like dopamine, serotonin, opioids, and acetylcholine. These proteins are considered important for understanding the mechanisms behind addiction and how the brain responds to addictive substances [102]. It is possible that hnRNP K, along with other proteins involved in alterations in gene expression and RNA processing, could contribute to the development of ASD.

Cadherin-3 (P-Cadherin) is a cell adhesion protein that belongs to the cadherin family playing a crucial role in cell-cell adhesion and tissue morphogenesis during embryonic development. Alterations in cadherin expression and function have been implicated in various neurological disorders, including ASD. Cadherins have been considered as possible targets for cognitive disorders such as ASD over the past years [103]. These findings suggest that P-Cadherin may contribute to the pathogenesis of ASD.

Due to the variabilities in the samples, and a limited sample size, developing a simple model that can detect the disorder is not an easy task. This work is a result of trying many preprocessing and FS techniques. Below a summary of all tried out methods is brought.

High leverage points (HLPs) are observations in LR that are made at extreme positions in the space of explanatory variables and are far from the average of the data [104]. HLPs can mask outliers and need to be detected and if necessary removed. We remove the HLPs from the dataset based on [104], and with a 5-fold LR, we reach a train AUC of 0.91, but a test AUC of 0.798. The limitation of this work is that HLPs are removed from the whole dataset and the test samples are adjusted. Moreover, the labels are considered when finding the HLPs. This makes the analysis not trustworthy.

This dataset has outliers and robust methods are needed to detect them and limit their effect on the analysis. In [45], an HRT is suggested which consists of two stages, the Quick Rejection Stage (QRS) and the Precise Rejection Stage (PRS) to filter data before training the diagnostic model and remove outliers. However, this method finds the outliers in each group separately using its statistics which could lead to the exclusion of valid data points.

In other part of our study, we fit a linear and polynomial functions to the cases and controls, and use their coefficients as variables. However, since this dataset has HLPs (they could also be outliers or low quality samples due to an error), fitting a robust and accurate function is not possible.

Another challenge of this dataset is that the variables are heavily skewed. Box-cox transformation and winsorizing are tested to change the distribution of the data. However, in the end, we decide that only adjusting the z-scores to fall in the range of (-3, 3) without further manipulating the data is enough.

For FS, many studies select the significant variables by evaluating their t-statistics. We use Student's t-test, and moderated t-test presented in [105], and Bonferroni correction to p-values. However, many of the features pass the significance test and therefore, we cannot limit the feature space.

Because of significant variations among individuals due to confounding factors such as age, gender, diet, and comorbid diseases, many protein levels within the same group vary so significantly that they cannot be used as reliable biomarkers. Therefore, reproducibility and robustness of data mining algorithms for disease detection through proteomic profiling of patient

serum have yet to be established. This work seeks to analyze the proteomic data from a systems engineering perspective, i.e., applying principles and adapting proven techniques for fault detection and diagnosis in system engineering to ASD detection. From a systems engineering point of view, sensor measurements in chemical processes are highly correlated because of the network structure and physical and chemical principles that govern the process operation. Serum protein concentrations are correlated for biological systems such as the human body because of biophysical and biochemical principles. Based on the system level similarities between a chemical plant and a human body, this paper postulates that the correlations among some proteins in the serum are changed accordingly. We aim to identify significant correlation changes instead of specific protein level changes to tackle the ASD detection problem. Since the correlations among different proteins are tightly controlled by gene regulatory networks (GRNs) to ensure biological functions, the within-group variations of these correlations are expected to be much smaller than the variations of the individual protein levels. The correlation changes specifically caused by the disease can be detected and used as alternative biomarkers for ASD detection. With this idea, we propose to find pairs of proteins that have either high correlation in one group and the other does not, or have high correlation in both groups but with opposite direction, or both have low correlation but have a difference in their mean and the pair shows separation. However, our analysis shows that in a 2D feature space, finding a pair of feature based on spearman's rank correlation coefficient is not possible and most of the samples overlap and separation is not done.

Different filtering and wrapper FS techniques are tested out in this study. None of them performs well alone and due to this, we try the hybrid FS technique to take advantages of both methods. Some of the filtering methods tested are reversed correlation algorithm (RCA) [106], correlation-based feature selection (CFS) [107], and MIT correlation, which is also known as signal-to-noise statistic [108]. CFS's results are better than other methods however, it does not improve the performance of the model from the baseline (AUC=0.86 in [43]). A few of wrapper methods utilized are ReliefF [109], and RFE which overfit. Due to the curse of dimentionality, RFE starts with a model that is overfitting already and cannot help finding accurate features.

3.6 Conclusion

ASD is a complex and heterogeneous disorder influenced by various genetic, environmental, and biological factors. While previous studies suggest a promising role for the most of the biomarkers found by our algorithm, multiple genes and biological pathways are involved in the development of ASD. Therefore, further research is needed to fully understand the involvement of these proteins in ASD detection, and their significance in the broader context of ASD etiology. Longitudinal studies and larger-scale investigations are necessary to validate the diagnostic accuracy and clinical utility of the suggested biomarkers.

We introduce a novel set of engineered features, incorporating protein ratios, which mitigate within-class variations by their resilience to confounding factors. Additionally, our FS technique integrates a sequential filter and wrapper method to address their individual constraints, followed by the development of a linear ML model. This model is chosen for its resistance to overfitting and superior interpretability.

The proposed methodology introduces systems engineering principles and techniques to provide new insights into early ASD detection research. Biomarkers beyond the traditional physical trait are defined to include bioinformation extracted from proteomic data or any other system-level measurements. The systems engineering perspective for disease detection will enable future research to obtain valuable insights on the ASD mechanism and, as such, eventually leads to additional discoveries.

Conducting pathway enrichment analysis on the selected biomarkers can help ensure their accuracy by identifying their functions and significant pathways, thereby enhancing the effectiveness of this study. Moreover, employing a completely separate dataset for testing the model's performance can emphasize the efficacy of the proposed framework.

Chapter 4

Detecting Pulmonary Arterial Hypertension: A Case Study Using the Biomarker Detection Framework from chapter 3

Pulmonary arterial hypertension (PAH) is a severe complication frequently associated with SSc, affecting a notable percentage of individuals with this condition. PAH significantly threatens patient survival, with right heart failure being a leading cause of death in this patient group. Early detection of PAH is crucial due to its significant impact on patient outcomes, as individuals with SSc and PAH experience markedly reduced survival rates compared to those without PAH. However, the diverse clinical presentations and lack of specific symptoms in the early stages often result in delayed diagnosis. Various organizations have issued screening recommendations, primarily relying on symptoms and echocardiography findings. However, current methods have limitations, prompting the exploration of new screening approaches leveraging blood biomarkers data. Recent studies have also explored novel diagnostic biomarkers, such as the insulin-like growth factor (IGF) axis and mitogen-activated protein kinase 6 (MAPK6), enhancing the understanding of PAH molecular mechanisms and potential therapeutic targets. Biomarkers beyond the traditional physical trait are needed to include bioinformation extracted from proteomic data or any other system-level measurements. Since the algorithm proposed in chapter 3 shows promise as a groundbreaking tool for early disease detection for ASD, we aim to assess its effectiveness by examining the performance of a selected panel of biomarkers for PAH detection, comparing it with established models to determine its potential impact. Results suggest that with 8 biomarkers we can improve AUC of the PAH detection model by 9.8-16.5% based on different ML models.

4.1 Introduction

PAH is a severe complication frequently associated with SSc, impacting approximately 7–12% of individuals with this condition [110, 111]. It poses a significant risk to patient survival, as right heart failure, a common consequence of PAH, stands as the leading cause of death in this patient group, accounting for approximately 26% of fatalities [112]. Individuals with SSc and PAH have a markedly reduced survival rate, estimated at 56% over three years compared to 94% for those without PAH [113]. In Europe, SSc-PAH represents 15–20% of all PAH cases with 30% mortality rate within the first year [114, 115, 116].

The diagnosis of PAH typically involves a mean pulmonary arterial pressure (mPAP) of 25 mm Hg or higher, coupled with a pulmonary capillary wedge pressure (PCWP) of 15 mm Hg or lower. Unfortunately, due to the diverse clinical presentations and the lack of specific symptoms in the early stages, diagnosis is often delayed. Early detection of PAH is crucial, as observational studies suggest that identifying and intervening in the disease's early stages can significantly improve patient outcomes [117]. Therefore, there's a pressing need for improved screening methods to detect PAH early and expedite diagnosis and intervention [118].

Various organizations, including the American College of Cardiology Foundation/American Heart Association and the European Society of Cardiology/European Respiratory Society, have issued screening recommendations, predominantly relying on symptoms and echocardiography findings [119]. However, current methods exhibit limitations, such as variable application and the inability to capture all patients accurately [117]. New screening approaches leverage blood biomarkers, imaging techniques, and real-world healthcare data [118]. The DETECT algorithm, integrating biomarkers like N-terminal pro-brain natriuretic peptide (NT-proBNP) and uric acid, has shown promise in early PAH detection, outperforming traditional methods [117]. Additionally, recent studies have explored novel diagnostic biomarkers, including the IGF axis [120] and MAPK6 [121], shedding light on potential therapeutic targets and enhancing the understanding of the molecular mechanisms underlying PAH. In one study, a panel of 6 biomarker is discovered to detect PAH with an AUC of 0.866 [118].

The algorithm proposed in chapter 3 shows promise as a groundbreaking tool for early disease detection across various disorders/diseases. In this study, we aim to assess its effectiveness by examining the performance of a selected panel of biomarkers for PAH detection. Using the Sheffield confirmatory cohort, we compare the algorithm's performance with that of the model developed in [118] to determine its potential impact.

4.2 Materials and Methods

4.2.1 Dataset

An cohort was selected from treatment-naive PAH patients in The Sheffield Teaching Hospitals Observational Study (STH-ObS) of Patients with Pulmonary Hypertension (PH), Cardiovascular, and Lung Disease. These patients underwent right heart catheterization (RHC) at the Sheffield Pulmonary Vascular Disease Unit (Royal Hallamshire Hospital, Sheffield, UK), following approval from the Research Ethics Committee 18/YH/0441. Serum samples were collected from diagnostic RHC procedures conducted between 2008 and 2015 before the formal diagnosis of PH. The cohort consisted of individuals with a confirmed diagnosis of SSc and evidence indicative of PAH but without concurrent interstitial lung disease (PAH n=23). Additionally, a disease control group was established, comprising SSc patients with confirmed diagnoses but negative RHC findings for PH (non-PH n=22). Serum samples from all participants were stored at -80°C until analysis [118].

4.2.2 Preprocessing

In this study, a total of 296 protein analytes are initially detected. However, 8 of these analytes are excluded since they do not meet quality standards. Additionally, one PAH sample is removed from the analysis as it has 27 missing analytes. Subsequently, analytes with more than 50% missing values are excluded from further analysis. For the remaining 232 protein analytes, missing values are imputed using the KNN imputer from the *sklearn* library. Ratios are then generated from the protein levels, followed by a log2 transformation and z-scaling. To address outliers, z-transformed values below -3 and above 3 are adjusted to -3 and 3, respectively.

4.2.3 Methods

Novel features facilitate identifying biomarkers related to PH, necessitating feature engineering techniques. Additionally, we address the challenge of the curse of dimensionality by carefully selecting relevant features, recognizing the significance of this step in our analysis. We aim to validate the accuracy of the framework developed in the previous chapter by applying it to a dataset associated with another disease. If our framework demonstrates high performance in detecting PH, it underscores its novelty and potential applicability to various diseases. Therefore, the methods outlined in this section closely mirror those in the previous chapter, with the key difference being the fine-tuning of model hyperparameters to suit the characteristics of the new dataset.

4.3 Results

Using a specific random seed, our analysis identified eight features based on Figure 4.1: IL-6R beta/CFH, Vitronectin/MIF, VEGF-D/ErbB3, MIF/IL-1RII, VEGF-D/IL-1RII, Sclerostin, TIMP-1/CFH, KLK-7/IL-1RII with an AUC of 1.000 for both training and 0.929 test datasets with an LDA classifier. To further assess the robustness of these selected features, we train a LR, LDA, SVM, and GNB classifiers 20 times (20 MCVTs), calculating the AUC for each iteration. The results are presented in Table 4.1. The results suggest 9.8-16.5% improvement over the previous work on this dataset [118].

For a better comparison, we find the performance of the 6 proteins suggested in [118] with LDA, LR, SVM, and GNB. Results in Table 4.2 suggest that performance of the proposed features in this study not only result in higher AUC, but also they have more consistent performance and as a result smaller standard deviation across all models.

4.4 Discussion

In this section, we review literature to see if the features suggested by our algorithm has been correlated to PAH before.

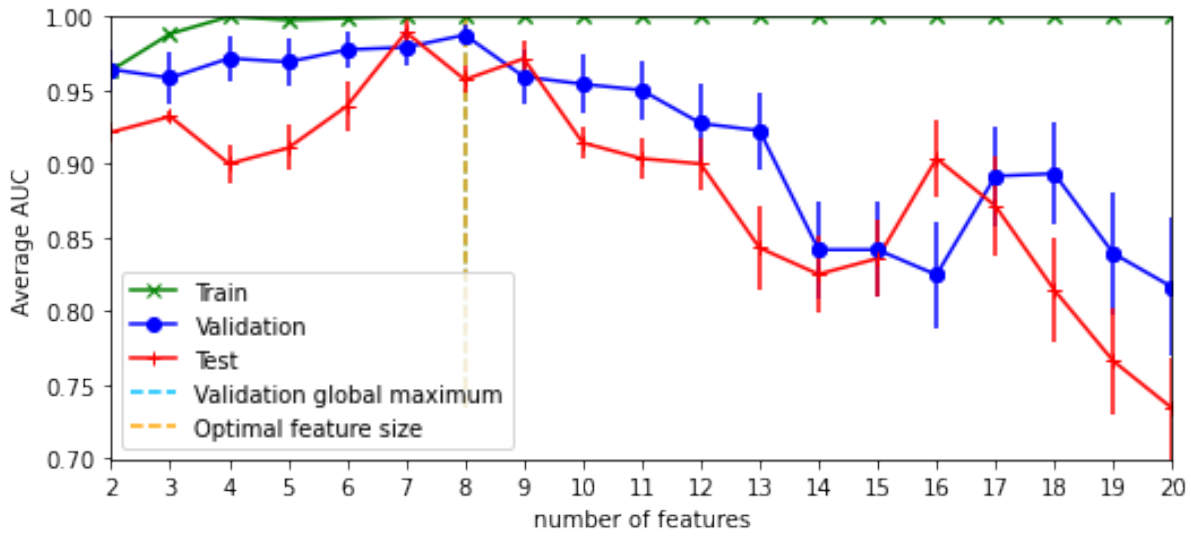

Figure 4.1: The plot of average AUC for a range of number of features. The selection is based on the AUC of validation (red plot). The global maximum of this curve is shown with a green dotted line. The optimal model size is shown with yellow dotted line.

Table 4.1: AUC of LR, LDA, SVM, and GNB classifiers on the panel of 8 biomarkers proposed by our algorithm

MCVT	**LR**	**LDA**	**SVM**	**GNB**
1	1.000	1.000	1.000	1.000
2	1.000	1.000	1.000	1.000
3	1.000	1.000	1.000	1.000
4	1.000	1.000	1.000	1.000
5	1.000	1.000	1.000	1.000
6	1.000	1.000	0.950	0.750
7	1.000	0.900	0.000	1.000
8	0.950	0.900	1.000	0.700
9	0.833	0.833	0.889	0.833
10	0.950	1.000	0.100	0.950
11	1.000	1.000	1.000	1.000
12	0.944	0.889	0.944	0.944
13	1.000	0.929	0.929	0.857
14	1.000	0.900	1.000	1.000
15	1.000	1.000	0.000	1.000
16	0.900	0.900	0.900	0.950
17	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000
mean	0.979	0.963	0.836	0.949
std	0.044	0.055	0.348	0.091

Table 4.2: AUC of LR, LDA, SVM, and GNB classifiers on the panel of 6 proteins proposed by [118]

MCVT	LR	LDA	SVM	GNB
1	0.950	0.850	0.850	0.900
2	1.000	1.000	1.000	1.000
3	1.000	0.900	1.000	1.000
4	1.000	1.000	1.000	1.000
5	0.650	0.750	0.700	0.700
6	1.000	1.000	0.000	1.000
7	0.750	0.750	0.250	0.750
8	0.700	0.750	0.700	0.650
9	0.778	0.722	0.889	0.722
10	0.950	0.950	0.200	0.900
11	1.000	1.000	1.000	1.000
12	1.000	1.000	1.000	1.000
13	1.000	1.000	0.929	1.000
14	0.800	0.700	0.800	0.800
15	0.944	0.833	0.944	1.000
16	0.750	0.750	0.750	0.750
17	0.800	0.800	0.700	0.850
18	1.000	1.000	0.000	1.000
19	0.833	0.778	0.833	0.833
20	0.929	0.857	0.786	0.929
mean	0.892	0.870	0.717	0.889
std	0.120	0.115	0.331	0.122

Interleukin-6 receptor subunit beta (IL-6R beta) is a protein involved in the signaling pathway of interleukin-6 (IL-6), a cytokine that plays various roles in inflammation, immune response, and hematopoiesis. In a study, it is suggested that IL-6 is elevated in the serum of patients with PH and is negatively correlated with lung function in those patients. Additionally, it suggests that IL-6 is one of the most important mediators in the pathogenesis of inflammation in PH [122]. Since IL-6R beta is a part of the receptor complex through which IL-6 exerts its effects, it is plausible that IL-6R beta could be implicated in the pathogenesis of PH as well. Further studies directly linking IL-6R beta to PH is necessary to confirm this hypothesis.

Complement Factor H (CFH) is a regulatory protein involved in the complement system, which is part of the immune system and plays a role in inflammation, immunity, and tissue homeostasis. In a study high-throughput analysis of the plasma proteome in PAH patients revealed decreases in the inhibitor CFH. This decrease in CFH, along with increases in the

activator of the alternative pathway, CFD, was associated with a high risk of mortality in PAH patients [123].

Vitronectin is a glycoprotein involved in various physiological processes, including cell adhesion, wound healing, and regulation of the complement system. A study reports a decrease in plasminogen activator inhibitor-1 (PAI-1) expression in patients with idiopathic pulmonary arterial hypertension (IPAH) compared to healthy donors. Additionally, the study suggests that decreased PAI-1 levels in IPAH may lead to increased vitronectin levels, potentially contributing to pathologic pulmonary vascular remodeling [124].

Macrophage Migration Inhibitory Factor (MIF) is a pro-inflammatory cytokine known to play a role in various inflammatory and autoimmune conditions. Elevated MIF levels correlate with disease severity in PH/PAH, suggesting its potential as a biomarker [125].

Vascular endothelial growth factor D (VEGF-D) protein plays a crucial role in promoting lymphangiogenesis and angiogenesis, contributing to the formation of new lymphatic vessels and blood vessels in the body. One study suggests that in the Sugen 5416/chronic hypoxia (SuHx) rat model of severe angioobliterative PAH, there is increased expression of VEGF-D in the lungs, which may contribute to the development of pulmonary vascular disease [126].

Receptor tyrosine-protein kinase erbB-3 (ErbB3) protein is a member of the epidermal growth factor receptor (EGFR) family, involved in various cellular processes, including cell proliferation, survival, and differentiation. While there may not be direct studies on ErbB3 and PAH, alterations in signaling pathways involving growth factors like those mediated by ErbB3 could potentially contribute to the pathogenesis of PAH. Further research may be needed to explore the specific role of ErbB3 in pulmonary vascular function and its potential involvement in PAH.

Interleukin-1 receptor type 2 (IL-1RII) is a decoy receptor that binds to IL-1 cytokines, regulating their signaling and modulating inflammatory responses in the body. While there is no direct study on how this protein can be affected by PAH, IL-1 cytokines, which interact with IL-1RII, have been implicated in various inflammatory and immune-related diseases, such as PH [127].

Sclerostin is a protein that inhibits osteoblastic bone formation, contributing to the regulation of bone mineralization. One research suggested a positive correlation between serum sclerostin levels and PH in pre-dialysis end-stage kidney disease (ESKD) patients, indicating that higher sclerostin levels were associated with echocardiographic structural cardiac abnormalities, particularly PH, in the patient population [128].

Tissue Inhibitor of Metalloproteinases 1 (TIMP-1) is a protein that helps regulate the activity of enzymes called metalloproteinases, which are involved in tissue remodeling and repair processes in the body. In one study, it is indicated that up-regulation of histone deacetylases 1 (HDAC1) contributes to the development of PAH which in the end influences the levels of TIMP-1 [129].

Kallikrein-7 (KLK-7) is a serine protease involved in various physiological processes, including skin desquamation and inflammatory responses. Although there is not a study for the direct effect of PAH on KLK-7, one study suggested that increased concentrations of KLK-1, KLK-3, KLK-7, KLK-8, and KLK-12 may have a strong correlation with conditions like hypertension, inflammation, and obesity [130]. Further research is needed to find the effect of PAH on KLK-7 levels.

4.5 Conclusion

PAH often complicates SSc, posing a significant threat to affected individuals. PAH-related right heart failure is a leading cause of mortality in this patient population, emphasizing the importance of early detection to improve outcomes. However, diagnosing PAH in SSc patients is challenging due to varied clinical presentations and lack of specific symptoms, leading to delayed recognition. While current screening methods rely on symptoms and echocardiography, their limitations have prompted exploration of novel approaches incorporating blood biomarkers. Incorporating biomarkers derived from proteomic data or other system-level measurements is essential for enhancing disease detection beyond traditional physical traits.

Given the promising nature of the algorithm proposed in chapter 3 for early disease detection in ASD, evaluating its efficacy in detecting PAH using a selected panel of biomarkers compared to established models is crucial. Initial results suggest a 10-16% enhancement in

performance by leveraging protein ratios and systematically meaningful features, indicating the potential impact of this approach.

This study has its own limitations including low sample size. Despite this limitation, we optimize hyperparameters in model training to minimize standard deviation in results, enhancing the reliability of identified biomarkers. Additionally, the biomarkers suggested by this study should be assessed using bioinformatics methods such as pathway analysis. Moreover, validation of the suggested biomarkers using an independent dataset is essential to confirm their robustness and applicability across diverse populations.

Chapter 5

Summary and Future Work

This chapter summarizes the contributions of this dissertation and discusses the limitations and potential future work.

In this work, I aim to advance the detection of a type of disorder in body by proposing novel ML-assisted frameworks. These frameworks consist of preprocessing, feature engineering, feature selection, model training, and performance evaluation. The integration of domain knowledge with the ML, we engineer physically or statistically meaningful features. The feature selection methods help identify the relevant and predictive features for classification of the two groups of controls and cases. This research, through different projects, shows that ML models integrated with knowledge-based features enhance the performance of detection and interpretability.

Chapter 2 aims to develop accurate and efficient screening methods for childhood speech disorders, ultimately facilitating early intervention and improved outcomes for patients. We introduce an automated speech disorders screening method for children, using their recordings uttering the word "flower", showcasing novel knowledge-based features' efficacy in detection. Our proposed features, including n-gram count and strength and ratio-based metrics, offer improved characterization of speech recordings. To address sample imbalances, we utilize SMOTE and synthesize minority samples. We present a two-step feature selection procedure to enhance classification accuracy and reduce model complexity. Comparative analysis demonstrates significant performance enhancements with selected features across linear and nonlinear classification methods, highlighting their effectiveness in characterizing speech disorders. Despite small sample sizes, careful model training and testing procedures, including MCVT and

FS, enable robust conclusions. The study findings emphasize the importance of investigating longer audio recordings and understanding the effectiveness of specific triage words like "flower," prompting further research and inclusion of similar words to enhance screening accuracy.

Chapter 3 aims to detect ASD using serum biomarkers with an automated method to facilitate the early detection. This study explores biomarkers' roles in detection and aims to validate biomarker diagnostic accuracy and their clinical use. Our study introduces novel engineered features, including protein ratios, to mitigate within-class variations and incorporates a sequential filter and wrapper feature selection method to develop a linear ML model, chosen for its resistance to overfitting and superior interpretability. By integrating systems engineering principles, our methodology offers fresh perspectives on early ASD detection, defining biomarkers beyond physical traits to encompass proteomic and system-level data. This approach promises valuable insights into ASD mechanisms and potential breakthroughs in future research endeavors. This work can benefit from pathway enrichment analysis for the optimal biomarkers to ensure the accuracy of the selected features by finding their functions and significant pathways. Employing a completely separate dataset for testing the model's performance can emphasize the efficacy of the proposed framework.

Chapter 4 aims to validate the ASD detection framework proposed in the previous chapter by applying it to another disorder, PAH. Many existing biomarker detection methods are limited to the training dataset, prompting our investigation into the effectiveness of the model for PAH detection using protein ratios and systematically meaningful features. Results indicate that the framework shows promise in detecting the disorder using blood biomarkers. Despite limitations such as low sample size, optimizing hyperparameters in model training enhances biomarker reliability. Furthermore, assessing suggested biomarkers using bioinformatics methods like pathway analysis and validating them with an independent dataset are essential steps to confirm their robustness and applicability.

References

[1] Patricia A Prelock, Tiffany Hutchins, and Frances P Glascoe. “Speech-language impairment: how to identify the most common and least diagnosed disability of childhood”. In: *The Medscape Journal of Medicine* 10.6 (2008), p. 136 (cit. on p. 3).

[2] Charles E Irwin et al. “Preventive care for adolescents: few get visits and fewer get services”. In: *Pediatrics* 123.4 (2009), e565–e572 (cit. on p. 3).

[3] American Academy of Pediatrics et al. “Council on children with disabilities, section on developmental behavioral pediatrics, bright futures steering committee, medical home initiatives for children with special needs project advisory committee. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening”. In: *Pediatrics* 118.1 (2006), pp. 405–420 (cit. on p. 3).

[4] Ramesh Raghavan et al. “Speech and language disorders in children: Implications for the Social Security Administration’s Supplemental Security Income Program”. In: (2016) (cit. on p. 3).

[5] Heidi D Nelson et al. “Screening for speech and language delay in preschool children: systematic evidence review for the US Preventive Services Task Force”. In: *Pediatrics* 117.2 (2006), e298–e319 (cit. on p. 3).

[6] Keiko Ishikawa, Joel MacAuslan, and Suzanne Boyce. “Toward clinical application of landmark-based speech analysis: Landmark expression in normal adult speech”. In: *The Journal of the Acoustical Society of America* 142.5 (2017), EL441–EL447 (cit. on pp. 4, 5).

[7] Ann A Tyler and Leslie C Tolbert. “Speech-language assessment in the clinical setting”. In: (2002) (cit. on p. 4).

[8] Carol Stoel-Gammon. “Transcibing the speech of young children”. In: *Topics in language disorders* 21.4 (2001), pp. 12–21 (cit. on p. 4).

[9] Martin J Ball and Joan Rahilly. “Transcribing disordered speech: The segmental and prosodic layers”. In: *Clinical linguistics & phonetics* 16.5 (2002), pp. 329–344 (cit. on p. 4).

[10] Visar Berisha, Rene Utianski, and Julie Liss. “Towards a clinical tool for automatic intelligibility assessment”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 2825–2828 (cit. on p. 4).

[11] Kenneth N Stevens. “Toward a model for lexical access based on acoustic landmarks and distinctive features”. In: *The Journal of the Acoustical Society of America* 111.4 (2002), pp. 1872–1891 (cit. on p. 4).

[12] Sharlene A Liu. “Landmark detection for distinctive feature-based speech recognition”. In: *The Journal of the Acoustical Society of America* 100.5 (1996), pp. 3417–3430 (cit. on p. 4).

[13] Andrew Wilson Howitt. “Automatic syllable detection for vowel landmarks”. In: (2000) (cit. on p. 4).

[14] John HL Hansen and Sanjay Patil. “Speech under stress: Analysis, modeling and recognition”. In: *Speaker classification I*. Springer, 2007, pp. 108–137 (cit. on p. 4).

[15] Zhaocheng Huang, Julien Epps, and Dale Joachim. “Investigation of speech landmark patterns for depression detection”. In: *IEEE Transactions on Affective Computing* (2019) (cit. on pp. 4–6).

[16] Keshi Dai, Harriet J Fell, and Joel MacAuslan. “Recognizing emotion in speech using neural networks”. In: *Telehealth and Assistive Technologies* 31 (2008), pp. 38–43 (cit. on p. 4).

[17] Tin Lay Nwe, Haizhou Li, and Minghui Dong. “Analysis and detection of speech under sleep deprivation”. In: *Ninth International Conference on Spoken Language Processing*. 2006 (cit. on pp. 4, 5).

[18] Marisha Speights Atkins et al. “Computer-assisted Syllable Complexity Analysis of Continuous Speech as a Measure of Child Speech Disorders”. In: *Proceedings of the 19th International Congress of Phonetic Sciences,(ICPhS 2019), Melbourne, Australia*. 2019, pp. 4–10 (cit. on p. 5).

[19] J MacAuslan et al. “Automated tools for identifying syllabic lndmark clusters that reflect changes in articulation”. In: *Automated Tools for Identifying Syllabic Lndmark Clusters that Reflect Changes in Articulation* (2011), pp. 63–66 (cit. on p. 5).

[20] Karen Chenausky, Joel MacAuslan, and Richard Goldhor. “Acoustic analysis of PD speech”. In: *Parkinson’s Disease* 2011 (2011) (cit. on p. 5).

[21] Chi-youn Park. “Consonant landmark detection for speech recognition”. PhD thesis. Massachusetts Institute of Technology, 2008 (cit. on p. 5).

[22] Zhaocheng Huang, Julien Epps, and Dale Joachim. “Speech landmark bigrams for depression detection from naturalistic smartphone speech”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 5856–5860 (cit. on p. 5).

[23] Hernandez-Diaz Huici et al. “Speech rate estimation in disordered speech based on spectral landmark detection”. In: *Biomedical signal processing and control* 27 (2016), pp. 1–6 (cit. on p. 6).

[24] Andrea Carolina Trevino, Thomas Francis Quatieri, and Nicolas Malyska. “Phonologically-based biomarkers for major depressive disorder”. In: *EURASIP Journal on Advances in Signal Processing* 2011.1 (2011), pp. 1–18 (cit. on p. 6).

[25] Marisha Speights Atkins, Dallin J Bailey, and Suzanne E Boyce. “Speech exemplar and evaluation database (SEED) for clinical training in articulatory phonetics and speech science”. In: *Clinical linguistics & phonetics* 34.9 (2020), pp. 878–886 (cit. on p. 7).

[26] Elisabeth H Wiig, Wayne A Secord, and Eleanor Semel. *CELF-Preschool-2: Clinical evaluation of language fundamentals, preschool*. Harcourt Assessment, 2004 (cit. on p. 7).

[27] Barbara Dodd et al. *Diagnostic evaluation of articulation and phonology (DEAP)*. Psychology Corporation, 2002 (cit. on p. 7).

[28] Warwick Williams et al. "The practicality of using a smart phone 'App'as an SLM and personal noise exposure meter (SoundLog)". In: *Proceedings of ACOUSTICS*. 2016, pp. 1–7 (cit. on p. 8).

[29] Carolyn Anderson and Wendy Cohen. "Measuring word complexity in speech screening: single-word sampling to identify phonological delay/disorder in preschool children". In: *International journal of language & communication disorders* 47.5 (2012), pp. 534–541 (cit. on pp. 8, 17).

[30] Nathalie Japkowicz and Shaju Stephen. "The class imbalance problem: A systematic study". In: *Intelligent data analysis* 6.5 (2002), pp. 429–449 (cit. on p. 11).

[31] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357 (cit. on pp. 11, 12).

[32] Guillaume Lemaıtre, Fernando Nogueira, and Christos K Aridas. "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 559–563 (cit. on p. 12).

[33] Devarshi Shah, Jin Wang, and Q Peter He. "A feature-based soft sensor for spectroscopic data analysis". In: *Journal of Process Control* 78 (2019), pp. 98–107 (cit. on pp. 12, 20, 35).

[34] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830 (cit. on pp. 14, 49).

[35] Robert Tibshirani and Jerome H Friedman. *The elements of statistical learning [electronic resource]: data mining, inference, and prediction: with 200 full-color illustrations*. Vol. 9. Springer, 2001 (cit. on pp. 15, 17, 52).

[36] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013 (cit. on pp. 16, 17, 51, 52).

[37] KW Lau and QH Wu. "Online training of support vector classifier". In: *Pattern Recognition* 36.8 (2003), pp. 1913–1920 (cit. on pp. 16, 17, 51, 52).

[38] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: a library for support vector machines". In: *ACM transactions on intelligent systems and technology (TIST)* 2.3 (2011), pp. 1–27 (cit. on p. 17).

[39] L Breiman. *Random Forests Mach Learn. 2001; 45: 5–32* (cit. on p. 17).

[40] Tianqi Chen and Carlos Guestrin. "Xgboost: A scalable tree boosting system". In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016, pp. 785–794 (cit. on p. 17).

[41] American Psychiatric Association et al. "Diagnostic and statistical manual of mental disorders (p. 317)". In: *Washington: American Psychiatric Association* (1997) (cit. on p. 29).

[42] Matthew J Maenner et al. "Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, United States, 2018". In: *MMWR Surveillance Summaries* 70.11 (2021), p. 1 (cit. on p. 29).

[43] Laura Hewitson et al. "Blood biomarker discovery for autism spectrum disorder: A proteomic analysis". In: *PLoS One* 16.2 (2021), e0246581 (cit. on pp. 29, 33–35, 41, 49, 55, 56, 58–64, 69).

[44] Joseph Abraham, Nicholas Szoko, and Marvin R Natowicz. "Proteomic investigations of autism spectrum disorder: past findings, current challenges, and future prospects". In: *Reviews on Biomarker Studies in Psychiatric and Neurodegenerative Disorders* (2019), pp. 235–252 (cit. on p. 29).

[45] Ahmed I Saleh and Asmaa H Rabie. "A new Autism Spectrum Disorder Discovery (ASDD) strategy using data mining techniques based on blood tests". In: *Biomedical Signal Processing and Control* 81 (2023), p. 104419 (cit. on pp. 29, 33, 62, 63, 68).

[46] Geraldine Dawson and Kathleen Zanolli. “Early intervention and brain plasticity in autism”. In: *Autism: Neural Basis and Treatment Possibilities: Novartis Foundation Symposium 251*. Vol. 251. Wiley Online Library. 2003, pp. 266–280 (cit. on p. 29).

[47] Afaf El-Ansary et al. “Preliminary evaluation of a novel nine-biomarker profile for the prediction of autism spectrum disorder”. In: *PLoS One* 15.1 (2020), e0227626 (cit. on p. 29).

[48] Safa Al-Amrani et al. “Proteomics: Concepts and applications in human medicine”. In: *World Journal of Biological Chemistry* 12.5 (2021), p. 57 (cit. on p. 29).

[49] Kayce H Ryberg. “Evidence for the implementation of the Early Start Denver Model for young children with autism spectrum disorder”. In: *Journal of the American Psychiatric Nurses Association* 21.5 (2015), pp. 327–337 (cit. on p. 29).

[50] Fatir Qureshi et al. “Multivariate Analysis of Metabolomic and Nutritional Profiles among Children with Autism Spectrum Disorder”. In: *Journal of Personalized Medicine* 12.6 (2022), p. 923 (cit. on pp. 29, 32, 33).

[51] Mei Sze Tan et al. “A review on omics-based biomarkers discovery for Alzheimer’s disease from the bioinformatics perspectives: statistical approach vs machine learning approach”. In: *Computers in biology and medicine* 139 (2021), p. 104947 (cit. on pp. 30, 31).

[52] Lihua Li et al. “Data mining techniques for cancer detection using serum proteomic profiling”. In: *Artificial intelligence in medicine* 32.2 (2004), pp. 71–83 (cit. on p. 30).

[53] T Mary-Huard, F Picard, and S Robin. “Introduction to statistical methods for microarray data analysis”. In: *Mathematical and Computational Methods in Biology* (2006), pp. 56–126 (cit. on p. 30).

[54] Aizatul Shafiqah Mohd Faizal et al. “A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach”. In: *Computer methods and programs in biomedicine* 207 (2021), p. 106190 (cit. on p. 31).

[55] Shan V Andrews et al. “Case-control meta-analysis of blood DNA methylation and autism spectrum disorder”. In: *Molecular autism* 9.1 (2018), pp. 1–11 (cit. on p. 31).

[56] H Ij. “Statistics versus machine learning”. In: *Nat Methods* 15.4 (2018), p. 233 (cit. on p. 31).

[57] Eleftherios P Diamandis. “Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems”. In: *Journal of the National Cancer Institute* 96.5 (2004), pp. 353–356 (cit. on p. 31).

[58] Keith A Baggerly et al. “Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer”. In: *Journal of the National Cancer Institute* 97.4 (2005), pp. 307–309 (cit. on p. 31).

[59] Kerul Suthar et al. “Feature engineering and machine learning for computer-assisted screening of children with speech disorders”. In: *PLOS Digital Health* 1.5 (2022), e0000041 (cit. on pp. 31, 35).

[60] Farnaz Yousefi Zowj et al. “Process Systems Engineering Guided Machine Learning for Speech Disorder Screening in Children”. In: *Computer Aided Chemical Engineering*. Vol. 49. Elsevier, 2022, pp. 1843–1848 (cit. on p. 31).

[61] Marta Lualdi and Mauro Fasano. “Statistical analysis of proteomics data: a review on feature selection”. In: *Journal of proteomics* 198 (2019), pp. 18–26 (cit. on p. 31).

[62] Suman Raj and Sarfaraz Masood. “Analysis and detection of autism spectrum disorder using machine learning techniques”. In: *Procedia Computer Science* 167 (2020), pp. 994–1004 (cit. on p. 32).

[63] Daniel P Howsmon et al. “Multivariate techniques enable a biochemical classification of children with autism spectrum disorder versus typically-developing peers: A comparison and validation study”. In: *Bioengineering & translational medicine* 3.2 (2018), pp. 156–165 (cit. on p. 32).

[64] Tania Akter et al. “Machine learning model to predict autism investigating eye-tracking dataset”. In: *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE. 2021, pp. 383–387 (cit. on p. 32).

[65] Tania Akter et al. “Improved machine learning based classification model for early autism detection”. In: *2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*. IEEE. 2021, pp. 742–747 (cit. on p. 32).

[66] Mehmet Baygin et al. “Automated ASD detection using hybrid deep lightweight features extracted from EEG signals”. In: *Computers in Biology and Medicine* 134 (2021), p. 104548 (cit. on p. 32).

[67] Juan Manuel Mayor Torres et al. “Facial emotions are accurately encoded in the neural signal of those with autism spectrum disorder: A deep learning approach”. In: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 7.7 (2022), pp. 688–695 (cit. on p. 32).

[68] Hossein Haghighat et al. “An age-dependent Connectivity-based computer aided diagnosis system for autism spectrum disorder using resting-state fMRI”. In: *Biomedical Signal Processing and Control* 71 (2022), p. 103108 (cit. on p. 32).

[69] Huan Liu and Lei Yu. “Toward integrating feature selection algorithms for classification and clustering”. In: *IEEE Transactions on knowledge and data engineering* 17.4 (2005), pp. 491–502 (cit. on p. 37).

[70] Isabelle Guyon and André Elisseeff. “An introduction to variable and feature selection”. In: *Journal of machine learning research* 3.Mar (2003), pp. 1157–1182 (cit. on p. 38).

[71] Robin Genuer, Jean-Michel Poggi, and Christine Tuleau-Malot. “Variable selection using random forests”. In: *Pattern recognition letters* 31.14 (2010), pp. 2225–2236 (cit. on p. 39).

[72] Yonghong Peng, Zhiqing Wu, and Jianmin Jiang. “A novel feature selection approach for biomedical data classification”. In: *Journal of Biomedical Informatics* 43.1 (2010), pp. 15–23 (cit. on pp. 39–41, 43–45).

[73] Aliasghar Shahrjooihaghighi et al. “An ensemble feature selection method for biomarker discovery”. In: *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE. 2017, pp. 416–421 (cit. on p. 39).

[74] Felipe Colombelli, Thayne Woycinck Kowalski, and Mariana Recamonde-Mendoza. "A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles". In: *Knowledge-Based Systems* 254 (2022), p. 109655 (cit. on p. 39).

[75] https://towardsdatascience.com/fishers-linear-discriminant-intuitively-explained-52a1ba79e1bb (cit. on p. 40).

[76] Alexander Statnikov et al. "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis". In: *Bioinformatics* 21.5 (2005), pp. 631–643 (cit. on p. 41).

[77] JM Linacre and G Rasch. "The expected value of a point-biserial (or similar) correlation". In: *Rasch Measurement Transactions* 22.1 (2008), p. 1154 (cit. on p. 44).

[78] Francesc J Ferri et al. "Comparative study of techniques for large-scale feature selection". In: *Machine intelligence and pattern recognition*. Vol. 16. Elsevier, 1994, pp. 403–413 (cit. on p. 45).

[79] Pavel Pudil, Jana Novovičová, and Josef Kittler. "Floating search methods in feature selection". In: *Pattern recognition letters* 15.11 (1994), pp. 1119–1125 (cit. on p. 46).

[80] Petr Somol et al. "Adaptive floating search methods in feature selection". In: *Pattern recognition letters* 20.11-13 (1999), pp. 1157–1163 (cit. on p. 46).

[81] Gareth James et al. *An introduction to statistical learning: With applications in python*. Springer Nature, 2023 (cit. on p. 48).

[82] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 49).

[83] David R Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 20.2 (1958), pp. 215–232 (cit. on p. 50).

[84] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009 (cit. on p. 52).

[85] Harry Zhang. “The optimality of naive Bayes”. In: *Aa* 1.2 (2004), p. 3 (cit. on p. 53).

[86] Thomas A Lasko et al. “The use of receiver operating characteristic curves in biomedical informatics”. In: *Journal of biomedical informatics* 38.5 (2005), pp. 404–415 (cit. on p. 54).

[87] John A Swets. “Measuring the accuracy of diagnostic systems”. In: *Science* 240.4857 (1988), pp. 1285–1293 (cit. on p. 54).

[88] Mark H Zweig and Gregory Campbell. “Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine”. In: *Clinical chemistry* 39.4 (1993), pp. 561–577 (cit. on p. 54).

[89] Frederic Mercier, Youngsu Cho Kwon, and Rich Kodama. “Meningeal/vascular alterations and loss of extracellular matrix in the neurogenic zone of adult BTBR T+ tf/J mice, animal model for autism”. In: *Neuroscience letters* 498.3 (2011), pp. 173–178 (cit. on p. 62).

[90] Elizabeth C Lee and Valerie W Hu. “Phenotypic subtyping and re-analysis of existing methylation data from autistic probands in simplex families reveal ASD subtype-associated differentially methylated genes and biological functions”. In: *International Journal of Molecular Sciences* 21.18 (2020), p. 6877 (cit. on p. 64).

[91] Juliana Magdalon et al. “Complement system in brain architecture and neurodevelopmental disorders”. In: *Frontiers in neuroscience* 14 (2020), p. 503589 (cit. on p. 64).

[92] Julie Ouellette. “Role of cerebrovascular abnormalities in the 16p11. 2 deletion autism syndrome”. PhD thesis. Université d’Ottawa/University of Ottawa, 2019 (cit. on p. 64).

[93] Abolghasem Tohidpour et al. “Neuroinflammation and infection: molecular mechanisms associated with dysfunction of neurovascular unit”. In: *Frontiers in cellular and infection microbiology* 7 (2017), p. 276 (cit. on p. 65).

[94] Elaine Y Hsiao et al. “Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders”. In: *Cell* 155.7 (2013), pp. 1451–1463 (cit. on p. 66).

[95] Valerio Napolioni et al. “Plasma cytokine profiling in sibling pairs discordant for autism spectrum disorder”. In: *Journal of neuroinflammation* 10 (2013), pp. 1–12 (cit. on p. 66).

[96] Jing Zhang et al. “AI Machine Learning Technique Characterizes Potential Markers of Depression in Two Animal Models of Depression”. In: *Brain Sciences* 13.5 (2023), p. 763 (cit. on p. 66).

[97] Elham Mousavinejad et al. “Coenzyme Q10 supplementation reduces oxidative stress and decreases antioxidant enzyme activity in children with autism spectrum disorders”. In: *Psychiatry research* 265 (2018), pp. 62–69 (cit. on p. 66).

[98] Feng Gu, Ved Chauhan, and Abha Chauhan. “Impaired synthesis and antioxidant defense of glutathione in the cerebellum of autistic subjects: alterations in the activities and protein expression of glutathione-related enzymes”. In: *Free Radical Biology and Medicine* 65 (2013), pp. 488–496 (cit. on p. 66).

[99] CB Lücking and A Brice*. “Alpha-synuclein and Parkinson’s disease”. In: *Cellular and Molecular Life Sciences CMLS* 57 (2000), pp. 1894–1908 (cit. on p. 67).

[100] Wilaiwan Sriwimol, Pornprot Limprasert, et al. “Significant changes in plasma alpha-synuclein and beta-synuclein levels in male children with autism spectrum disorder”. In: *BioMed research international* 2018 (2018) (cit. on p. 67).

[101] Christian Praml et al. “Genetic variation of Aflatoxin B1 aldehyde reductase genes (AFAR) in human tumour cells”. In: *Cancer letters* 272.1 (2008), pp. 160–166 (cit. on p. 67).

[102] Camron D Bryant and Neema Yazdani. “RNA-binding proteins, neural development and the addictions”. In: *Genes, Brain and Behavior* 15.1 (2016), pp. 169–186 (cit. on p. 67).

[103] Aziz El-Amraoui and Christine Petit. “Cadherins as targets for genetic diseases”. In: *Cold Spring Harbor perspectives in biology* 2.1 (2010), a003095 (cit. on p. 67).

[104] Anwar Fitrianto and Tham Wendy. “Identification of high leverage points in binary logistic regression”. In: *AIP Conference Proceedings*. Vol. 1782. 1. AIP Publishing. 2016 (cit. on p. 68).

[105] Kai Kammers et al. “Detecting significant changes in protein abundance”. In: *EuPA open proteomics* 7 (2015), pp. 11–19 (cit. on p. 68).

[106] Agnieszka Wosiak, Danuta Zakrzewska, et al. “Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis”. In: *Complexity* 2018 (2018) (cit. on p. 69).

[107] Mark A Hall. “Correlation-based feature selection for machine learning”. PhD thesis. The University of Waikato, 1999 (cit. on p. 69).

[108] Ying Liu. “Serum proteomic pattern analysis for early cancer detection”. In: *Technology in cancer research & treatment* 5.1 (2006), pp. 61–66 (cit. on p. 69).

[109] Igor Kononenko. “Estimating attributes: Analysis and extensions of RELIEF”. In: *European conference on machine learning*. Springer. 1994, pp. 171–182 (cit. on p. 69).

[110] D Mukerjee et al. “Prevalence and outcome in systemic sclerosis associated pulmonary arterial hypertension: application of a registry approach”. In: *Annals of the rheumatic diseases* 62.11 (2003), pp. 1088–1093 (cit. on p. 72).

[111] Eric Hachulla et al. “Early detection of pulmonary arterial hypertension in systemic sclerosis: a French nationwide prospective multicenter study”. In: *Arthritis & Rheumatism* 52.12 (2005), pp. 3792–3800 (cit. on p. 72).

[112] Anthony J Tyndall et al. “Causes and risk factors for death in systemic sclerosis: a study from the EULAR Scleroderma Trials and Research (EUSTAR) database”. In: *Annals of the rheumatic diseases* 69.10 (2010), pp. 1809–1815 (cit. on p. 72).

[113] Eric Hachulla et al. “Risk factors for death and the 3-year survival of patients with systemic sclerosis: the French ItinerAIR-Sclerodermie study”. In: *Rheumatology* 48.3 (2009), pp. 304–308 (cit. on p. 72).

[114] Marc Humbert et al. “Pulmonary arterial hypertension in France: results from a national registry”. In: *American journal of respiratory and critical care medicine* 173.9 (2006), pp. 1023–1030 (cit. on p. 72).

[115] Lorinda Chung et al. *Survival and predictors of mortality in systemic sclerosis-associated pulmonary arterial hypertension: outcomes from the pulmonary hypertension assessment and recognition of outcomes in scleroderma registry*. 2014 (cit. on p. 72).

[116] J Hurdman et al. “ASPIRE registry: assessing the Spectrum of Pulmonary hypertension Identified at a REferral centre”. In: *European Respiratory Journal* 39.4 (2012), pp. 945–955 (cit. on p. 72).

[117] J Gerry Coghlan et al. “Evidence-based detection of pulmonary arterial hypertension in systemic sclerosis: the DETECT study”. In: *Annals of the rheumatic diseases* 73.7 (2014), pp. 1340–1349 (cit. on p. 72).

[118] Yasmina Bauer et al. “Identifying early pulmonary arterial hypertension biomarkers in systemic sclerosis: machine learning on proteomics from the DETECT cohort”. In: *European Respiratory Journal* 57.6 (2021) (cit. on pp. 72–74, 76).

[119] David B Badesch et al. “Diagnosis and assessment of pulmonary arterial hypertension”. In: *Journal of the American College of Cardiology* 54.1_Supplement_S (2009), S55–S66 (cit. on p. 72).

[120] Melanie K Nies et al. “Proteomics discovery of pulmonary hypertension biomarkers: insulin-like growth factor binding proteins are associated with disease severity”. In: *Pulmonary Circulation* 12.2 (2022), e12039 (cit. on p. 72).

[121] Xiaomei Yang et al. “Identification of crucial hub genes and differential T cell infiltration in idiopathic pulmonary arterial hypertension using bioinformatics strategies”. In: *Frontiers in Molecular Biosciences* 9 (2022), p. 800888 (cit. on p. 72).

[122] Wei-Jie Xu et al. “Interleukin-6 and pulmonary hypertension: from physiopathology to therapy”. In: *Frontiers in Immunology* 14 (2023), p. 1181987 (cit. on p. 76).

[123] Christopher J Rhodes et al. “Plasma proteome analysis in patients with pulmonary arterial hypertension: an observational cohort study”. In: *The Lancet Respiratory Medicine* 5.9 (2017), pp. 717–726 (cit. on p. 77).

[124] Fotini M Kouri et al. “Plasminogen activator inhibitor type 1 inhibits smooth muscle cell proliferation in pulmonary arterial hypertension”. In: *The International Journal of Biochemistry & Cell Biology* 40.9 (2008), pp. 1872–1882 (cit. on p. 77).

[125] Gael Jalce and Christophe Guignabert. “Multiple roles of macrophage migration inhibitory factor in pulmonary hypertension”. In: *American Journal of Physiology-Lung Cellular and Molecular Physiology* 318.1 (2020), pp. L1–L9 (cit. on p. 77).

[126] Ayser Al-Husseini et al. “Vascular endothelial growth factor receptor 3 signaling contributes to angioobliterative pulmonary hypertension”. In: *Pulmonary circulation* 5.1 (2015), pp. 101–116 (cit. on p. 77).

[127] David M Essayan et al. “Biologic activities of IL-1 and its role in human disease”. In: *Journal of allergy and clinical immunology* 102.3 (1998), pp. 344–350 (cit. on p. 77).

[128] Jonghyun Lee et al. “Higher sclerostin is associated with pulmonary hypertension in pre-dialysis end-stage kidney disease patients: a cross-sectional prospective observational cohort study”. In: *BMC Pulmonary Medicine* 24.1 (2024), p. 78 (cit. on p. 78).

[129] Fangwei Li et al. “Inhibition of HDAC1 alleviates monocrotaline-induced pulmonary arterial remodeling through up-regulation of miR-34a”. In: *Respiratory Research* 22.1 (2021), p. 239 (cit. on p. 78).

[130] Kang Li et al. “Transcriptome reveals the overexpression of a kallikrein gene cluster (KLK1/3/7/8/12) in the Tibetans with high altitude-associated polycythemia”. In: *International Journal of Molecular Medicine* 39.2 (2017), pp. 287–296 (cit. on p. 78).