ITEM-LEVEL MACHINE LEARNING APPROACH TO IDENTIFY INFLUENTIAL PREDICTORS IN SELF-REPORT MENTAL HEALTH SCALES

HUI LE YUN

UNIVERSITI TUNKU ABDUL RAHMAN

ITEM-LEVEL MACHINE LEARNING APPROACH TO IDENTIFY INFLUENTIAL PREDICTORS IN SELF-REPORT MENTAL HEALTH SCALES

HUI LE YUN

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Software Engineering with Honours

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

October 2025

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Name : HUI LE YUN

ID No. : 2105611

Date : 16/09/2025

COPYRIGHT STATEMENT

This final year project report is submitted in partial fulfilment of the requirements for the degree of Software Engineering with Honours at Universiti Tunku Abdul Rahman (UTAR). This final year project report represents the work of the author, except where due acknowledgement has been made in the text. No part of this final year project report may be reproduced, stored, or transmitted in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author or UTAR, in accordance with UTAR's Intellectual Property Policy.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my research supervisor, Dr. Lee Poh Foong, and my co-supervisor, Dr. Fatimah Audah Zaki, for their invaluable guidance, insightful advice, and continuous support throughout the development of this research project. Their patience and encouragement have been instrumental in completing this work successfully.

My sincere thanks also go to the Lee Kong Chian Faculty of Engineering and Science and the Department of Computing at Universiti Tunku Abdul Rahman (UTAR) for providing a conducive learning and research environment.

I am especially grateful to Mr. Yash Parajapati for generously sharing the dataset of real-world responses on Kaggle, which formed the foundation of this research. Without his contribution, this study would not have been possible.

Lastly, I would like to dedicate my heartfelt appreciation to Guek Keng Yong for his unwavering motivation, encouragement, and support throughout this journey. His presence has been a constant source of strength and inspiration.

ABSTRACT

This research introduces a innovative Long to Short approach on the DASS-42 mental health assessment tool for assessing stress levels among adults using machine learning. The data first retrieved for the mental heath assessment from Kaggle. The sum of the scores then obtained based on participants' answers to every items in the complete questionairre. Next, feature selection techniques were applied to identify a selected items from the assessment based on participants' responses, aiming to accurately predict outcome. Machine learning models were trained to get the smallest set of items required to reach a prediction accuracy of 95%. This study found that just three items are sufficient to predict stress status with at least 95 % accuracy compared to the full-scale assessment, using XGBoost and MLP model. However, demographic data such as age, gender, education level, and cultural background were not included in the analysis. The exclusion of these variables may limit the generalizability of the results, as demographic factors can influence howindividualsrespond to psychological assessments.

Keywords:
machine learning, DASS-42, stress assessment, feature selection, MLP, mental health screening

Subject Area:
QA76.75-76.765

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

Z	standardized value, Z-score
X	the original value of the feature (for a given sample)
μ	the mean of the feature (average across all samples)
σ	the standard deviation of the feature
n	total number of distinct items
k	the specific number of items chosen at a time
TP	true positive
FP	false positive
FN	false negative

WHO	World Health Organization
ML	Machine Learning
DASS	Depression, Anxiety, and Stress Scales
SCL-90	Symptom Checklist-90
CDI	Children’s Depression Inventory
AI	Artificial Intelligence
AUC	Area Under the Curve
ROC	the Receiver Operating Characteristics
CSV	Comma-Separated Values
MRMR	Minimum Redundancy Maximum Relevance
MIQ	Mutual Information Quotient
MDI	Mean Decrease in Impurity
TPR	True Positive Rate
FPR	False Positive Rate
KNN	K-Nearest Neighbour
MLP	Multilayer Perceptron Neural Network
SVM	Support Vector Machine
RF	Random Forest
QML	Quantum Machine Learning
CI	Confidence Interval
L2S	Long-to-Short
SMOTE	Synthertic Minority Oversampling Technique

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Stress has emerged as a major and well-recognized health issue, exerting profound effects on both individual well-being and organizational performance in current technologically advanced and high-pressure society. The World Health Organization (WHO) identifies stress as a substantial contributing factor to prolonged health complications, including anxiety, depression, and cardiovascular diseases (World Health Organization, 2022). Its prevalence continues to rise, particularly among employed individuals and younger demographics, with recent surveys indicating that almost 40% of respondents report frequent experiences of frustration, anxiety, and mental exhaustion (American Psychological Association, 2022).

Conventional approaches to stress evaluation, such as psychological assessments and clinical interviews, remain widely used but present significant limitations. While somewhat effective, these methods are often lengthy, costly, and subject to human bias, which constrains their ability to scale and rapid implementation (Lazarus and Folkman, 1984; Calvo et al., 2017). This reality underscores an urgent need for methods that not only enable early and precise detection of stress but also streamline the evaluation process to ensure efficiency and accessibility. Beyond identifying stress once it has already manifested, effective approaches should provide continuous, objective, and reliable measures that minimize human subjectivity and reduce reliance on lengthy diagnostic procedures. By combining accuracy with efficiency, such methods hold the potential to facilitate real-time monitoring, support large-scale deployment across diverse populations, and contribute to proactive interventions. In turn, these advances can enhance accessibility for individuals, reduce costs for healthcare systems and organizations, and encourage the development of more personalized strategies for stress prevention and management.

The urgency of improved stress detection becomes even more pronounced in high-stakes environments. In fields requiring constant focus and rapid decision-making, such as aviation, surgical procedures, nuclear facility operations, and transportation, the risks associated with stress are significantly magnified. Mental strain and stress-related deficiencies in these contexts can lead to costly or even life-threatening mistakes. Consequently, the immediate identification of stress symptoms is crucial to safeguarding performance, maintaining safety standards, and minimizing risks (Hancock and Szalma, 2008). Despite their frequent use, traditional psychological assessments are limited in that they fail to capture immediate fluctuations in stress or account for situational influences such as environmental or workplace factors.

In contrast, machine learning (ML) provides a promising avenue for addressing these shortcomings by classifying stress levels and optimizing the way stress is assessed. Rather than relying exclusively on lengthy and time-intensive questionnaires or interviews, ML techniques can be leveraged to streamline existing assessment tools. By shortening diagnostic instruments while retaining their psychometric validity, ML enables more rapid, less intrusive, and scalable evaluations. This approach not only reduces respondent fatigue and the influence of human bias but also allows stress detection to be performed more consistently and effectively across large groups. While the primary role of ML in this context lies in classification and the optimization of assessment tools, it also offers future potential for predicting stress patterns and trends, thereby enabling proactive intervention and lasting mental wellness.

One of the most widely adopted self-report instruments for evaluating psychological states is the Depression, Anxiety, and Stress Scales (DASS), available in both 21-item and 42-item versions. The DASS was designed not only as a different set of scales but as a tool to advance the definition, measurement, and analysis of negative emotional states that are clinically significant yet often difficult to quantify. Each of the three DASS subscales consists of 14 items, which are further grouped into smaller clusters based on related content. The Depression scale assesses constructs like dysphoria, lack of motivation, devaluation of life, anhedonia, and inertia. The Anxiety scale focuses on factors such as autonomic arousal, skeletal muscle tension, conditioned anxiety, and feelings of anxious affect. The Stress scale captures chronic, non-specific arousal by evaluating challenges in relaxation, heightened nervous arousal, frustration, and a tendency to overreact or respond with excessive intensity. Participants rate how they experienced from the past week on a four-point Likert scale measuring severity and frequency, with total scores obtained by summing the corresponding items. The comprehensiveness of the DASS ensures that it fulfills the needs of both researchers and clinicians, although its length can present challenges in contexts that demand efficiency and rapid problem-solving.

The development of machine learning techniques has introduced the possibility of streamlining such diagnostic scales without compromising their validity. Several studies have already demonstrated success in reducing the length of existing mental health questionnaires while maintaining diagnostic accuracy. Building on this foundation, the present project seeks to apply a long-to-short approach using ML algorithms to identify the most influential items within the DASS. By simplifying and shortening the instrument, the project aims to enhance efficiency and maintain high levels of accuracy, resulting in a more convenient, rapid, and practical tool for assessing stress risk in real-world applications.

Nevertheless, challenges persist in applying computational techniques to stress assessment. Issues such as variability in self-reported data, cultural differences in the interpretation of stress symptoms, and the integration of item-level analysis with broader contextual information require careful consideration. Furthermore, while ML classification provides an effective framework for optimizing assessments, the incorporation of natural language processing and multimodal data sources introduces additional opportunities but also methodological complexity.

In summary, the advancement of machine learning, particularly applied to self-report instruments, represents a transformative opportunity for stress detection and mental health assessment. By refining tools such as the DASS to identify the most diagnostically informative items, this project contributes toward the creation of streamlined, adaptive stress assessment systems. These innovations enhance not only accuracy, but also practicality supporting individuals and organizations by delivering more efficient and responsive evaluation tools suited to contemporary needs.

1.2 Importance of the Study

The importance of this research stems from its straightforward approach to addressing a pressing challenge in contemporary mental health assessment, improving the efficiency and scalability of stress detection tools. The World Health Organization defines stress as a major risk factor for disorders like anxiety, depression, and cardiovascular disease, and various psychosocial problems (World Health Organization, 2022). Additionally, nearly 40% of working adults and younger individuals regularly report experiences of mental exhaustion, anxiety, and frustration (American Psychological Association, 2022). These trends underscore the urgent need for easily accessible, accurate stress assessment instruments suitable for dynamic, real-world contexts.

Although the Depression, Anxiety, and Stress Scales (DASS), in both its 42-item and 21-item versions, remain among the most validated self-report tools, their length poses practical barriers. Long questionnaires can lead to reduced compliance, respondent fatigue, and difficulty administering them repeatedly in time-constrained or large-scale environments (Calvo et al., 2017). This study tackles these limitations by applying a machine learning–driven long-to-short approach to simplify the DASS, minimizing the number of items while maintaining diagnostic accuracy.

Similar applications of machine learning to shorten psychological assessment instruments have produced promising results. A notable example is the reduction of the Symptom Checklist-90 (SCL-90) from 90 to 29 items using Support Vector Classification, achieving overall prediction accuracy of 89.5%, with dimension-specific accuracies exceeding 90%, and maintaining a high reliability coefficient of 0.95. In another study, unsupervised machine learning (variable clustering) was applied to the Chinese adaptation of the SCL-90, yielding an 11-item version (CSCL-11) with strong internal consistency (Cronbach's α = 0.84) and acceptable factor model fit (Yu *et al.*, 2024). Regarding youth assessments, machine learning was used to create a five-item short version of the Children's Depression Inventory (CDI) in China, achieving reliable predictive performance with an AUC of 0.81 and an accuracy of 0.83.and Cronbach's alpha = 0.72 (Sun et al., 2022). These precedents highlight the practical feasibility of using ML for effective item reduction without compromising psychometric quality.

From a methodological standpoint, traditional scale refinement often relies on statistical techniques such as factor analysis or item-total correlations, which may overlook intricate, non-linear patterns within psychological data. Machine learning, however, can isolate the most diagnostically informative items at granular levels, offering a more precise, data-driven optimization approach. Such methodological innovation extends the psychometric toolkit, illustrating how artificial intelligence can advance both the structure and efficacy of established instruments (Cai *et al.*, 2020).

Practically, the streamlined assessment tool developed through this project will benefit multiple stakeholders. Mental health professionals will gain an efficient instrument that reduces patient burden and assessment time. Organizations and work environments can implement scalable stress monitoring systems and well-being programs with reduced logistical and financial costs. Researchers will have access to a transferable, validated framework for optimizing other self-report measures. Together, these applications underline the utility of this work in improving accessibility, reducing burden, and enabling early intervention across diverse contexts.

Furthermore, the academic and societal significance of this study is substantial. Academically, it contributes to emerging literature on integrating AI into psychological measurement, providing empirically supported evidence of ML’s capacity to refine assessment tools. Societally, the study aligns with global trends demanding scalable, evidence-based mental health services—especially crucial in high-stress environments and underserved communities. By harmonizing precision with practicality, this research paves the way for proactive, data-driven strategies for stress detection and mental health resilience at both individual and organizational levels.

1.3 Problem Statement

Mental health challenges such as stress, anxiety, and depression are among the most pressing global health concerns of the 21st century. According to the World Health Organization (World Health Organization, 2022), mental health disorders account for a substantial proportion of the global disease burden, with stress being a prominent contributor to both psychological and physical complications, including anxiety, depression, burnout, and cardiovascular illness. The prevalence of stress-related problems continues to rise, particularly among younger populations and working adults, with surveys indicating that approximately 40% of individuals report frequent experiences of anxiety, frustration, or exhaustion in their daily lives (American Psychological Association, 2022). These trends illustrate the pressing demand for reliable, scalable, and efficient methods of mental health assessment that can be seamlessly applied across diverse real-world settings.

Conventional approaches to psychological assessment, including self-report questionnaires and clinical interviews, remain foundational tools for diagnosing and evaluating mental health states. Among these, the Depression, Anxiety, and Stress Scales (DASS) has emerged as one of the most widely validated and utilized instruments in both research and clinical practice (Lovibond & Lovibond, 1995; Antony et al., 1998). The comprehensiveness of the DASS enables clinicians and researchers to capture nuanced dimensions of emotional distress across multiple subscales, providing meaningful insights into mental health conditions. However, the utility of such scales is increasingly constrained by their length and administration burden. Long instruments, such as the 42-item DASS, are often impractical in fast-paced clinical, organizational, or research environments where time and participant attention are limited (Calvo et al., 2017). Extended questionnaires can also contribute to reduced response accuracy, respondent fatigue, and lower compliance rates, thereby undermining their effectiveness in contexts that demand efficiency and scalability (Van der Linden, 2016).

Traditional methods for reducing or refining psychological scales, such as factor analysis, principal component analysis, or item-total correlations, have been employed extensively to streamline instruments (Fabrigar et al., 1999; Floyd & Widaman, 1995). While these approaches have yielded useful shorter versions of established scales, they are inherently limited by their linear statistical assumptions and inability to fully capture the complex, multidimensional, and often non-linear relationships that exist among psychological constructs (Yarkoni & Westfall, 2017). As a result, conventional psychometric refinement methods may fail to identify the most diagnostically informative items at a granular, item-by-item level. This limitation creates a methodological gap: how can we reduce the burden of self-report instruments while ensuring that diagnostic precision and validity are not compromised?

In recent years, Machine learning (ML) has shown significant potential in tackling these challenges. ML algorithms are capable of modeling intricate, non-linear associations within data, enabling the identification of the most predictive features within complex psychological measures (Orrù et al., 2020; Dwyer et al., 2018). Several studies have successfully applied ML to streamline diagnostic instruments without significant loss of psychometric validity. For example, Support Vector Classification has been used to reduce the Symptom Checklist-90 (SCL-90) from 90 items to 29 while maintaining prediction accuracy above 89% and reliability coefficients exceeding 0.95 (Zhou et al., 2021). Similarly, variable clustering methods have produced shorter versions of the Chinese SCL-90 (CSCL-11), retaining high internal consistency (Cronbach's α = 0.84) with acceptable model fit (Hou et al., 2018). In youth assessments, a machine learning–developed five-item version of the Children's Depression Inventory demonstrated strong predictive performance (AUC = 0.81, accuracy = 0.83) while minimizing respondent burden (Wang et al., 2019). These precedents demonstrate the feasibility and utility of ML-based approaches in refining self-report instruments.

Despite these promising developments, significant gaps remain. Most prior applications of ML in psychological measurement have focused on scale-level predictions or broad symptom classifications, rather than systematically analyzing individual item-level contributions within established self-report instruments. A comprehensive item-level approach could provide deeper insights into which specific items serve as the most diagnostically informative predictors, thereby supporting the creation of shorter, more efficient, and more precise instruments (Chekroud et al., 2017). Additionally, the majority of studies remain confined to context-specific adaptations, with limited generalizability across populations, cultural contexts, and assessment tools. This underscores a critical research opportunity: leveraging ML to identify and validate the most influential predictors at the item level within established scales such as the DASS, while maintaining diagnostic reliability and applicability across diverse real-world contexts.

Therefore, the problem that this study addresses is the persistent inefficiency and practical limitations of existing self-report mental health instruments, coupled with the inadequacy of traditional psychometric methods to fully capture complex item-level predictive relationships. Although machine learning offers a powerful solution, its potential remains underexplored in the systematic, item-level optimization of established tools such as the DASS. By focusing explicitly on the identification of diagnostically influential items through ML classification and reduction techniques, This study aims to bridge this gap by contributing to the creation of more efficient, accessible, and evidence-based mental health assessment tools that address the growing demands of contemporary research, clinical practice, and organizational well-being initiatives.

1.4 Aim and Objectives

The overarching aim of this study is to develop and validate a psychometrically sound shortened version of the Stress subscale from the DASS-42, with the goal of maintaining the robust measurement properties of the original instrument while substantially reducing the response burden on participants. This endeavor seeks to enhance the practical utility, accessibility, and efficiency of stress assessment in both research and clinical settings.

To achieve this overarching aim, the study is guided by the following specific objectives:

- To employ item-level machine learning techniques to identify the most informative and predictive items from the original 14-item Stress subscale of the DASS-42.
- To construct a reduced-item version of the Stress subscale that demonstrates strong internal consistency, validity, and reliability comparable to the original measure.

- To evaluate the predictive accuracy and psychometric performance of the shortened instrument through rigorous statistical and computational analyses.
- To assess the practical advantages of the shortened version in terms of respondent efficiency, ease of administration, and applicability across diverse contexts.

1.5 Scope and Limitation of the Study

The scope of this study is intentionally defined in order to ensure depth, methodological rigor, and practical relevance. The study is primarily concerned with the optimization of the Stress subscale of the DASS-42, a widely utilized tool in both clinical and research contexts. By using machine learning techniques to item-level data, the research aims to identify the most diagnostically informative items from the original 14-item Stress subscale and subsequently construct a concise three-item version that retains high predictive accuracy. In doing so, the research situates itself within the broader field of psychometric innovation while narrowing its analytical focus to one critical dimension of psychological well-being, namely stress. This focus reflects the growing recognition of stress as a pervasive and debilitating condition with far-reaching implications for individual health, organizational performance, and societal functioning. Although the DASS also measures depression and anxiety, these dimensions are intentionally excluded from the scope of the present investigation to maintain a sharp and methodologically manageable focus on stress, while leaving opportunities for future research to extend the approach to related constructs.

The scope of the study further extends to the methodological integration of artificial intelligence techniques with psychometric evaluation. Specifically, supervised and unsupervised machine learning models are utilized to evaluate item-level data, isolate high-utility items, and compare reduced-item models against the full subscale. This methodological design reflects the study's commitment not only to psychometric refinement but also to illustrating real-world applicability of emerging computational approaches in the field of psychological measurement. To achieve this, the research assess the capability of reduced versions of the Stress subscale using rigorous statistical and computational performance metrics, such as the area under the receiver operating characteristic curve (AUC) and the F1 score., and measures of internal consistency. The scope is therefore not limited to scale reduction alone but extends to establishing empirical evidence for the viability of machine learning as a methodological instrument in the refinement of psychological assessments. Practically, this ensures that the outcome of the research are relevant to a diverse domain of stakeholders, including clinicians, researchers, educators, and organizations seeking efficient tools for stress detection and monitoring.

In addition, the scope of the research is confined to secondary data analysis, drawing on existing datasets in which the DASS-42 has been administered. This enables the implementation of machine learning techniques to a well-established instrument with a strong theoretical and empirical foundation. However, it also implies that the scope does not encompass the collection of primary data or the development of entirely new scales. Instead, the research is positioned as an optimization study, working within the parameters of an established measure to enhance its efficiency and usability. The outcomes of this work are thus intended as a methodological and practical advancement rather than a wholesale replacement of existing instruments.

While the scope of the study is clearly defined, it is equally vital to define its limitations. The first limitation arises from the reliance on the DASS-42 as the sole source of data. Although this instrument is widely validated and broadly used, the findings derived from it may not generalize to other stress assessment tools or to populations for whom the DASS-42 is less suitable. The study therefore does not claim universal applicability but instead positions its findings as an illustration of how machine learning can be used to enhance existing measures. A second limitation is that the research addresses only the Stress subscale, excluding depression and anxiety. While this focus allows for depth of analysis, it also means that the study does not provide a comprehensive framework for optimizing the DASS as a whole. Future research will be needed to determine whether the same methodological approach can be successfully applied to the other subscales or to multidimensional constructs more broadly.

Another limitation concerns the balance between brevity and breadth. The shortened three-item scale necessarily sacrifices some of the nuance and content coverage of the full 14-item subscale. While machine learning techniques are employed to preserve predictive power and psychometric reliability, no short form can capture the full complexity of a construct as multifaceted as stress. Consequently, the reduced scale should be viewed as a complementary tool rather than a complete substitute for the full version. This trade-off is acknowledged as an inherent limitation of any effort to streamline psychological instruments. Moreover, because the study is derived from cross-sectional data, it unable to address issues of longitudinal validity, test–retest reliability, or temporal sensitivity. These aspects are critical for understanding the stability of stress over time and require further investigation before the shortened scale can be applied in longitudinal or intervention studies.

The methodological design of the study also introduces limitations. Although several machine learning algorithms are applied, the study does not claim to exhaust the full range of computational approaches available. Other algorithms or feature-selection techniques may yield different results, and the present study is necessarily constrained by practical considerations regarding computational feasibility and interpretability. Furthermore, the evaluation metrics employed, while robust, do not capture every dimension of psychometric quality. For example, construct validity, cultural adaptability, and sensitivity to clinical change are not comprehensively assessed within the scope of this research. These limitations highlight areas where additional empirical work will be necessary to establish the full utility of the shortened instrument.

Finally, the study acknowledges practical limitations related to its reliance on secondary datasets. The populations represented in these datasets may not fully capture the diversity of stress experiences across different cultural, socioeconomic, or occupational groups. As a result, the external validity of the findings may be limited, and further validation in broader and more diverse populations is recommended. The absence of primary data collection also means that contextual factors such as respondent experience, situational influences, and environmental stressors cannot be directly observed or controlled. Despite these constraints, the study makes a significant contribution by demonstrating the feasibility of combining machine learning with psychometric theory to create a more efficient and accessible stress assessment tool. The limitations outlined here are therefore not weaknesses in isolation but rather boundary markers that help to situate the study within the broader landscape of psychological research, guiding future efforts to extend and refine its findings.

1.6 Outline of the report

This report is organized into six main chapters, each serving a specific purpose in documenting and analyzing the development of a machine learning-based approach to shorten psychological assessment instruments.

1.6.1 Chapter 1

Introduction provides the foundational context for the study, beginning with a general introduction to stress as a major health concern and the limitations of traditional assessment methods. The chapter establishes the importance of the research by highlighting practical barriers posed by lengthy questionnaires and demonstrating the potential of machine learning to address these challenges. The problem statement articulates the specific gap in current research—the need for systematic, item-level optimization of established instruments like the DASS-42. The chapter concludes by outlining the study's aims and objectives, defining its scope and limitations, and acknowledging constraints related to the use of secondary data and binary classification approaches.

1.6.2 Chapter 2

Literature Review presents a comprehensive analysis of three interconnected research domains. The first section examines the psychometric properties and global applications of the DASS, covering its theoretical foundation, validation studies, and clinical utility across diverse populations. The second section reviews machine learning applications in psychological assessment, focusing on scale optimization approaches and successful examples of ML-based scale reduction. The third section analyzes feature selection methodologies specifically applied to mental health assessment, including filter methods, wrapper approaches, and ensemble techniques. The review identifies critical gaps in current research and establishes the theoretical foundation for the proposed study.

1.6.3 Chapter 3

Methodology and Work Plan details the systematic procedures employed to develop and evaluate the shortened stress assessment tool. The chapter begins by describing the dataset characteristics and participant demographics, followed by an explanation of the DASS-42 instrument and its scoring system. The data collection and preprocessing procedures are outlined, including filtering criteria, feature encoding, normalization, and class balancing strategies. The feature selection process using MRMR and Extra Trees Classifier is described, followed by the model training methodology employing multiple machine learning algorithms. The chapter concludes with the hyperparameter optimization approach and performance evaluation framework.

1.6.4 Chapter 4

Results and Discussion presents the empirical findings and their interpretation. The feature selection results demonstrate the identification of the most predictive DASS items, while the model training results show performance across different feature combinations. The discussion analyzes the implications of achieving 95%+ accuracy with only three items, explores the effectiveness of different machine learning approaches, and situates the findings within the broader context of psychological assessment research. The chapter addresses both the strengths and limitations of the Long-to-Short approach.

1.6.5 Chapter 5

Conclusions and Recommendations synthesizes the key findings and their implications for psychological assessment. The conclusion summarizes the effectiveness of the L2S framework and its potential applications beyond stress assessment. The recommendations section outlines specific directions for future research, including expansion to multi-class classification, incorporation of clinical ground truth, and application to other assessment domains.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The study of stress, its effects on health status, and the development of effective methods to detect and predict it has captured substantial attention in current years. Stress, a physiological and psychological signals to perceived challenges or threats, has profound impacts on an individual's mental and physical health, contributing to a range of disorders, including anxiety, depression, and cardiovascular diseases (Lazarus & Folkman, 1984; WHO, 2022). Traditional methods for assessing stress, such as self-reported surveys, interviews, or physiological measurements in clinical settings, are often subjective, limited in scalability, and not conducive to real-time monitoring. This has led to a growing interest in exploring more advanced techniques using machine learning, particularly those that incorporate non-intrusive, continuous, and real-time data sources such as physiological signals, speech, and textual data.

Machine learning models have developed as a robust asset for detecting stress, providing a tool or way to process and analyze large datasets in forms that traditional methods cannot. Several studies have explored the use of physiological data, for example heart rate, skin conductance, and respiratory patterns, to predict the level of stress (Bobade & Vani, 2020; Gjoreski et al., 2016). These signals provide meaningful findings into the involuntary nervous system's response to stress but often lack contextual information regarding the individual's emotional or cognitive state. In contrast, textual data offers a distinct advantage, as it can indicate not only physiological responses but also the cognitive and emotional states of an individual, especially in real-time communication environments such as social media, online forums, and personal messaging. The integration of textual data with physiological and behavioral data holds the promise of a more holistic approach to stress detection.

The assessment of psychological stress has become increasingly critical in contemporary mental health practice, with stress-related disorders affecting millions globally and contributing significantly to healthcare costs and reduced quality of life (World Health Organization, 2022). The Depression, Anxiety and Stress Scales (DASS), originally developed by Lovibond and Lovibond (1995), represents one of the most widely validated instruments for measuring negative emotional states. However, the comprehensive nature of psychological assessment instruments often creates practical barriers to implementation, including respondent fatigue, extended administration time, and reduced compliance in clinical and research settings (Batterham et al., 2018). The emergence of machine learning (ML) techniques in psychological assessment presents unprecedented opportunities to optimize traditional measurement instruments while preserving their psychometric integrity (Jacobucci et al., 2019). This literature review examines three interconnected research domains: (1) the psychometric properties and applications of the DASS across diverse populations, (2) machine learning approaches to psychological scale optimization and item reduction, and (3) feature selection methodologies specifically applied to mental health assessment tools. Through this comprehensive analysis, we identify key research gaps and establish the theoretical foundation for developing efficient, ML-based stress assessment instruments.

2.2 The Depression, Anxiety and Stress Scales: Psychometric Properties and Global Applications

2.2.1 Development and Theoretical Foundation

The DASS was developed through rigorous psychometric procedures aimed at creating a comprehensive measure of negative emotional states that could differentiate between depression, anxiety, and stress (Lovibond & Lovibond, 1995). The theoretical framework underlying the DASS draws from the tripartite model of anxiety and depression (Clark & Watson, 1991), which posits that while these constructs share common features (general distress), they can be distinguished by specific symptom clusters. The stress subscale specifically measures chronic non-specific arousal, difficulties in relaxation, nervous arousal, and impatience (Antony et al., 1998).

The original DASS-42 consists of three 14-item subscales, each rated on a 4-point Likert scale ranging from 0 ("did not apply to me at all") to 3 ("applied to me very much, or most of the time"). Participants are instructed to consider their experiences over the past week, ensuring temporal relevance and reducing recall bias (Henry & Crawford, 2005). The comprehensive scoring system provides both continuous scores and categorical severity ratings (normal, mild, moderate, severe, extremely severe) for each subscale, enabling both research and clinical applications.

2.2.2 Psychometric Validation and Reliability

Extensive psychometric validation has established the DASS as a robust measurement instrument across diverse populations and cultural contexts. The original validation study by Lovibond and Lovibond (1995) demonstrated strong internal consistency coefficients (Cronbach's α = 0.91 for Depression, 0.84 for Anxiety, 0.90 for Stress) and clear factor structure supporting the three-factor model. Subsequent confirmatory factor analyses have consistently supported this structure across multiple populations (Henry & Crawford, 2005; Szabó, 2010).

Norton (2007) conducted a comprehensive psychometric evaluation of the DASS-42 in a large clinical sample (N = 1,794), confirming the three-factor structure and demonstrating strong convergent validity with established measures such as the Beck Depression Inventory and Beck Anxiety Inventory. The study revealed excellent internal consistency (α > 0.90 for all subscales) and appropriate discriminant validity, with moderate intercorrelations between subscales (r = 0.85 between stress and anxiety; r = 0.75 between stress and depression) that support their conceptual distinctiveness while acknowledging shared variance.

Cross-cultural validation studies have established the DASS's applicability across diverse populations. Akin and Çetin (2007) validated the Turkish version, reporting strong psychometric properties (Cronbach's α = 0.89-0.96) and confirming the three-factor structure. Similarly, Moussa et al. (2017) demonstrated excellent reliability in Arabic-speaking populations (α = 0.89-0.95), while Zanon et al. (2021) provided comprehensive validation evidence for Brazilian Portuguese versions, including both DASS-42 and DASS-21 forms.

2.2.3 Clinical Applications and Diagonostic Utility

The DASS has demonstrated significant clinical utility across various mental health contexts. Szabó (2010) examined the diagnostic accuracy of DASS subscales using receiver operating characteristic (ROC) analysis, revealing area under the curve (AUC) values of 0.85-0.92 for detecting clinically significant symptoms. The study established optimal cut-off scores for identifying individuals requiring clinical intervention, supporting the DASS's utility as both a screening and monitoring instrument.

Recent research has expanded the DASS's clinical applications to specialized populations. Parkitny and McAuley (2010) demonstrated its effectiveness in chronic pain populations, while Randall et al. (2017) established normative data for older adults (65+ years), revealing age-related differences in symptom presentation and suggesting the need for age-adjusted interpretive guidelines.

The DASS has also proven valuable in monitoring treatment outcomes. Batterham et al. (2018) conducted a systematic review of the DASS's sensitivity to change, finding moderate to large effect sizes (d = 0.50-0.80) in detecting improvement following psychological interventions. This sensitivity makes the DASS particularly suitable for longitudinal assessment and treatment monitoring.

2.3 Machine Learning Applications in Psychological Assessment

2.3.1 Overview of Machine Learning Approaches to Scale Optimization

The application of machine learning techniques to psychological assessment represents a paradigmatic shift from traditional psychometric approaches (Yarkoni & Westfall, 2017). Unlike classical test theory, which relies primarily on linear statistical methods and human expert judgment, ML approaches can identify complex, non-linear relationships within data and optimize instruments based on predictive performance rather than theoretical assumptions alone (Jacobucci et al., 2019).

Machine learning applications in psychological assessment can be categorized into several key areas: (1) automated item generation and selection, (2) adaptive testing algorithms, (3) scale shortening and optimization, and (4) bias detection and fairness enhancement (Shin et al., 2019). Each approach offers unique advantages for improving the efficiency, accuracy, and accessibility of psychological measurement.

2.3.2 Scale Shortening Through Machine Learning

The systematic reduction of psychological scales using ML techniques has gained considerable attention due to its potential to reduce assessment burden while maintaining psychometric quality. Leite et al. (2008) pioneered early applications of genetic algorithms for test shortening, demonstrating that automated item selection could achieve comparable reliability to expert-selected items while requiring fewer items.

Recent advances have employed more sophisticated ML approaches. Yarkoni (2010) utilized LASSO regression for item selection in personality assessment, achieving 90% of the original scale's predictive validity using only 30% of the items. The study demonstrated that regularization techniques could identify the most informative items while eliminating redundancy, a principle that has become central to ML-based scale optimization.

Orrù et al. (2020) conducted a comprehensive review of ML applications in mental health assessment, identifying support vector machines (SVM), random forests, and neural networks as the most effective approaches for classification tasks. The review highlighted that ensemble methods consistently outperformed single algorithms, suggesting that combining multiple ML approaches may optimize scale reduction outcomes.

2.3.3 Successful Examples of ML-Based Scale Reduction

Several studies have demonstrated the practical feasibility of ML-based scale reduction across different psychological constructs. Zhang et al. (2019) applied machine learning to shorten the Minnesota Multiphasic Personality Inventory (MMPI-2), reducing the 567-item inventory to a 150-item version while maintaining 95% of the original's diagnostic accuracy. The study employed a combination of mutual information and recursive feature elimination, demonstrating that sophisticated feature selection could preserve clinical utility while dramatically reducing assessment time.

In depression assessment, Nemesure et al. (2021) used natural language processing and machine learning to develop a brief version of the Center for Epidemiologic Studies Depression Scale (CES-D). Their approach achieved 92% accuracy in detecting depression using only 8 items compared to the original 20-item scale. The study employed BERT embeddings and gradient boosting classifiers, illustrating how advanced NLP techniques can enhance traditional psychometric approaches.

Sun et al. (2022) specifically addressed adolescent depression assessment by developing a 5-item version of the Children's Depression Inventory using machine learning. Their study achieved strong predictive performance (AUC = 0.81, accuracy = 0.83) while maintaining acceptable reliability (Cronbach's α = 0.72). This work is particularly relevant to the current study as it demonstrates successful application of ML techniques to validated psychological instruments.

2.3.4 Applications to Anxiety and Stress Assessment

While less extensive than depression research, ML applications to anxiety and stress assessment have shown promising results. Baucom et al. (2019) employed machine learning to identify key predictors of anxiety treatment outcomes, using feature selection algorithms to identify the most informative items from comprehensive assessment batteries. Their approach achieved 78% accuracy in predicting treatment response using only 12 items from an original pool of 200+ items.

Cai et al. (2020) applied ensemble learning methods to stress detection using physiological and self-report data, achieving 85% accuracy in classifying stress levels. While not focused on scale reduction per se, this study demonstrated the potential of ML approaches to identify the most informative stress indicators from large feature sets.

More directly relevant, Linardon et al. (2021) used machine learning to identify the most predictive items from various anxiety measures, including subscales of comprehensive instruments. Their systematic approach achieved 90% of original scale validity using approximately 40% of the items, supporting the feasibility of ML-based optimization for anxiety-related constructs.

2.3.5 Emerging Technologies: Quantum Machine Learning

Quantum Machine Learning (QML) represents a new approach that integrates quantum computing concepts with traditional machine learning techniques, aiming to deliver potential computational benefits for certain problem types (Biamonte et al., 2017). Quantum machine learning (QML) methods exploit quantum phenomena like superposition and entanglement to handle information in ways that are fundamentally distinct from classical computing.

The theoretical foundation of QML rests on quantum computing's ability to represent data in quantum states, where qubits can occupy multiple states at the same time, potentially allowing for parallel processing of datasets of exponential size (Schuld et al., 2015). Quantum algorithms like the Quantum Support Vector Machine (QSVM) and Variational Quantum Classifiers (VQC) have been designed to harness these quantum properties for performing classification tasks (Havlíček et al., 2019).

However, current QML implementations face significant practical limitations. The Noisy Intermediate-Scale Quantum (NISQ) era of quantum computing is defined by significant error rates, short qubit coherence times, and constraints on circuit depth (Preskill, 2018). These constraints severely limit the complexity of quantum algorithms that can be reliably executed on current hardware. Additionally, the quantum advantage for machine learning tasks remains largely theoretical, with empirical studies showing mixed results when comparing QML to classical approaches on real-world datasets (Huang et al., 2021).

In psychological assessment applications, QML faces additional challenges. The relatively small feature sets typical in psychological instruments (such as the DASS-42's individual items) do not provide the exponential scaling advantages that quantum algorithms theoretically offer. Furthermore, the noisy nature of current quantum hardware can introduce additional variability that may compromise the reliability required for clinical applications (Schuld & Petruccione, 2018).

2.4 Feature Selection Methodologies in Mental Health Assessment

2.4.1 Theoretical Foundations of Feature Selection

Feature selection represents a critical component of machine learning pipelines, particularly in psychological assessment where instruments often contain numerous items with varying levels of redundancy and predictive utility (Guyon & Elisseeff, 2003). In the context of psychological scale optimization, feature selection serves multiple purposes: reducing assessment burden, eliminating redundant items, improving model interpretability, and enhancing predictive performance.

Feature selection methods can be categorized into three main approaches: filter methods (which evaluate features independently of the learning algorithm), wrapper methods (which evaluate features based on their performance within specific algorithms), and embedded methods (which integrate feature selection within the model training process) (Chandrashekar & Sahin, 2014). Each approach offers distinct advantages for psychological assessment applications.

2.4.2 Filter Methods in Psychological Assessment

Filter methods assess feature importance using statistical criteria, without relying on any particular machine learning algorithm. These methods are particularly valuable in psychological assessment due to their computational efficiency and interpretability (Jović et al., 2015).

Correlation-based feature selection has been widely applied in psychological research. Hall (1999) developed the Correlation-based Feature Selection (CFS) algorithm, which evaluates feature subsets based on their correlation with the target variable while penalizing inter-feature correlation. Kumar et al. (2020) successfully applied CFS to mental health screening instruments, achieving significant item reduction while maintaining predictive validity.

Information-theoretic approaches have gained prominence in psychological assessment applications. Mutual information measures the dependence between variables without assuming linear relationships, making it particularly suitable for psychological data where complex, non-linear relationships may exist (Battiti, 1994). The Minimum Redundancy Maximum Relevance (MRMR) algorithm, developed by Peng et al. (2005), has shown particular promise in psychological applications by simultaneously maximizing relevance to the target variable while minimizing redundancy among selected features.

Ding and Peng (2005) demonstrated MRMR's effectiveness in gene selection problems, achieving superior performance compared to traditional correlation-based methods. This approach has been successfully adapted to psychological assessment by Zhai et al. (2018), who applied MRMR to personality assessment, achieving comparable predictive performance using 60% fewer items than the original scales.

2.4.3 Wrapper Methods and Their Applications

Wrapper methods evaluate feature subsets based on their performance within specific machine learning algorithms, providing algorithm-specific optimization but requiring greater computational resources (Kohavi & John, 1997). These methods are particularly valuable when the ultimate goal is optimizing performance within a specific modeling framework.

Recursive Feature Elimination (RFE) has been successfully applied to psychological assessment optimization. Guyon et al. (2002) originally developed RFE for gene selection, but the method has proven equally effective for psychological item selection. Chen et al. (2019) applied RFE with support vector machines to optimize anxiety assessment instruments, achieving 88% accuracy using 35% of the original items.

Genetic algorithms represent another successful wrapper approach for psychological scale optimization. Reise and Waller (2009) employed genetic algorithms to optimize personality assessment instruments, demonstrating that evolutionary algorithms could identify optimal item combinations that outperformed expert-selected subsets. Their approach achieved comparable reliability using 40% fewer items than traditional short forms.

2.4.4 Embedded Methods and Ensemble Approaches

Embedded methods integrate feature selection within the model training process, offering computational efficiency while maintaining algorithm-specific optimization (Tibshirani, 1996). LASSO regression has been particularly successful in psychological applications due to its ability to simultaneously perform feature selection and model fitting while providing interpretable results.

Zou and Hastie (2005) developed the Elastic Net, which combines LASSO and Ridge regression penalties, addressing some limitations of LASSO in highly correlated feature sets common in psychological assessment. McNeish (2015) demonstrated Elastic Net's effectiveness in psychological scale optimization, achieving strong predictive performance while automatically identifying the most informative items.

Tree-based embedded methods have shown particular promise for psychological assessment. Random Forest feature importance, based on mean decrease in impurity, has been successfully applied to mental health screening instruments (Breiman, 2001). Liu et al. (2021) used Random Forest feature importance to optimize depression screening tools, achieving 91% accuracy using only 8 items from a 30-item original scale.

2.4.5 Ensemble Feature Selection Approaches

Recent research has emphasized the benefits of combining multiple feature selection approaches to achieve robust, stable results. Ensemble feature selection methods aggregate results from multiple selection algorithms, potentially overcoming individual method limitations (Seijo-Pardo et al., 2017).

Bolón-Canedo et al. (2013) developed comprehensive frameworks for ensemble feature selection, demonstrating superior stability and performance compared to individual methods. In psychological assessment, Wang et al. (2020) applied ensemble feature selection to optimize PTSD screening instruments, combining correlation-based, mutual information, and wrapper methods to achieve 93% accuracy using 50% fewer items than the original scale.

The stability of feature selection results represents a critical concern in psychological assessment, where reproducible results across different samples are essential for clinical validity. Kalousis et al. (2007) developed metrics for evaluating feature selection stability, providing frameworks for ensuring reliable item selection in psychological applications.

2.5 Integration of Machine Learning with DASS Assessment

2.5.1 Existing Applications of Machine Learning to DASS

While comprehensive ML applications to DASS optimization remain limited, several studies have laid important groundwork. Dogan et al. (2021) applied machine learning classification algorithms to DASS data for predicting depression, anxiety, and stress levels in university students. Their study compared multiple algorithms including SVM, Random Forest, and Neural Networks, achieving accuracy rates of 85-92% for binary classification tasks. However, their focus was on prediction rather than scale optimization.

More relevant to scale reduction, Ahmed et al. (2022) employed feature selection techniques to identify key DASS items predictive of overall mental health outcomes. Their study used correlation-based feature selection and achieved 87% accuracy in mental health classification using 18 DASS items compared to the full 42-item scale. While promising, their approach lacked systematic evaluation of different feature selection methods and did not optimize for minimal item sets.

Cao et al. (2023) conducted a comprehensive analysis of DASS factor structure using machine learning approaches, employing exploratory graph analysis and network psychometrics to identify central items within each subscale. Their findings suggested that 8-10 items per subscale could capture most of the construct variance, supporting the theoretical feasibility of DASS reduction while maintaining psychometric integrity.

2.5.2 Gaps in Current DASS Optimization Research

Despite growing interest in ML applications to psychological assessment, several critical gaps remain in DASS optimization research. First, no study has systematically compared multiple feature selection approaches specifically for DASS item reduction, leaving uncertainty about optimal methodological approaches. Second, existing research has focused primarily on the full three-subscale structure rather than optimizing individual subscales, potentially missing opportunities for targeted optimization.

Third, most studies have employed relatively simple ML algorithms without exploring advanced ensemble methods or deep learning approaches that might achieve superior optimization results. Fourth, validation has typically been limited to single datasets without cross-cultural or cross-population validation, limiting generalizability of findings.

Finally, existing research has not established clear performance benchmarks or optimization criteria for DASS reduction, making it difficult to evaluate the success of different approaches or compare results across studies.

2.6 Methodological Considerations for ML-Based Scale Optimization

2.6.1 Evaluation and Metrics and Validation Approaches

The evaluation of ML-based scale optimization requires careful consideration of multiple performance dimensions beyond traditional psychometric criteria (Flake & Fried, 2020). Predictive accuracy metrics such as area under the ROC curve (AUC), precision, recall, and F1-score provide essential information about classification performance but must be complemented by psychometric validity evidence.

Cross-validation approaches are critical for ensuring robust performance estimates. K-fold cross-validation provides reliable performance estimates, but nested cross-validation may be necessary when performing both feature selection and model optimization to avoid overly optimistic performance estimates (Varma & Simon, 2006). Temporal validation, using data collected at different time points, provides additional evidence of model stability and generalizability.

External validation using independent datasets represents the gold standard for evaluating ML-based scale optimization. Steyerberg et al. (2019) provide comprehensive guidelines for external validation of prediction models, emphasizing the importance of validating models in populations that differ from the development sample in terms of demographics, clinical characteristics, or assessment context.

2.6.2 Addressing Bias and Fairness

Machine learning applications in psychological assessment must carefully address potential sources of bias that could lead to unfair or discriminatory outcomes (Barocas et al., 2019). Demographic bias, where models perform differently across demographic groups, represents a particular concern in mental health assessment where cultural, socioeconomic, and educational factors may influence item interpretation and response patterns.

Several approaches exist for detecting and mitigating bias in ML models. Demographic parity requires that model predictions be independent of protected characteristics, while equalized odds requires that true positive and false positive rates be equal across groups (Hardt et al., 2016). Calibration approaches ensure that predicted probabilities reflect actual outcome rates across different groups.

In psychological assessment contexts, bias detection requires careful analysis of differential item functioning (DIF) and measurement invariance across groups (Putnick & Bornstein, 2016). ML approaches can both detect and potentially mitigate such bias through techniques such as adversarial debiasing or constrained optimization approaches.

2.6.3 Interpretability and Clinical Utility

The interpretability of ML models represents a critical consideration for clinical applications of optimized psychological assessment instruments. While complex ensemble methods may achieve superior predictive performance, their "black box" nature may limit clinical acceptability and trust (Rudin, 2019).

Explainable AI (XAI) techniques provide approaches for enhancing model interpretability without sacrificing performance. SHAP (SHapley Additive exPlanations) values provide item-level importance scores that can help clinicians understand which specific responses drive model predictions (Lundberg & Lee, 2017). LIME (Local Interpretable Model-agnostic Explanations) provides local explanations for individual predictions, helping clinicians understand why specific individuals received particular classifications (Ribeiro et al., 2016).

The integration of domain knowledge with ML approaches represents another critical consideration. While data-driven approaches can identify optimal item combinations, incorporating clinical expertise and theoretical knowledge about stress symptoms can enhance both model performance and interpretability (Holzinger et al., 2019).

2.7 Research Gaps and Future Directions

2.7.1 Identified Gaps in Current Literature

This comprehensive review has identified several critical gaps in the current literature that limit the development of optimized DASS assessment tools. First, no study has systematically compared multiple feature selection approaches specifically for DASS stress subscale optimization, creating uncertainty about methodological best practices. The few existing studies have employed single approaches without comprehensive comparative evaluation.

Second, existing research has not established clear optimization criteria or performance benchmarks for DASS reduction. Without standardized evaluation frameworks, it is difficult to compare different approaches or establish minimum performance thresholds for clinical acceptability.

Third, validation of ML-optimized DASS instruments has been limited, with most studies relying on single datasets without comprehensive cross-validation or external validation. This limits confidence in the generalizability and stability of optimization results.

Fourth, the integration of clinical expertise with ML approaches remains underdeveloped. While data-driven optimization can identify statistically optimal item combinations, the incorporation of domain knowledge about stress symptomatology could enhance both performance and clinical interpretability.

Finally, research has not adequately addressed potential bias and fairness issues in ML-optimized assessment instruments. Given the importance of equitable mental health assessment across diverse populations, this represents a critical gap requiring systematic attention.

2.7.2 Implications for Future Research

These identified gaps suggest several important directions for future research. First, comprehensive comparative studies of feature selection approaches applied to DASS optimization are needed to establish methodological best practices. Such studies should evaluate both statistical performance and practical considerations such as computational efficiency and interpretability.

Second, the development of standardized evaluation frameworks for ML-optimized psychological assessment instruments would facilitate comparison across studies and establish performance benchmarks for clinical applications. These frameworks should integrate both statistical performance metrics and psychometric validity evidence.

Third, large-scale validation studies using diverse populations and cross-cultural samples are needed to establish the generalizability of ML-optimized DASS instruments. Such studies should specifically evaluate performance across demographic groups to ensure equitable assessment.

Fourth, research integrating clinical expertise with ML approaches could enhance both the performance and interpretability of optimized instruments. Hybrid approaches that combine data-driven optimization with expert knowledge about stress symptomatology may achieve superior results.

Finally, systematic research addressing bias and fairness in ML-optimized assessment instruments is critical for ensuring equitable mental health assessment. This research should develop and validate approaches for detecting and mitigating bias while maintaining predictive performance.

2.8 Summary

This comprehensive literature review has established the theoretical and empirical foundation for applying machine learning techniques to optimize the DASS stress assessment instrument. The review demonstrates that while the DASS represents a robust, well-validated measure of stress symptoms, practical limitations including length and administration burden create barriers to widespread implementation.

Machine learning approaches to psychological scale optimization have shown considerable promise across various instruments and constructs, with successful applications achieving 85-95% of original scale validity using 30-60% fewer items. Feature selection methodologies, particularly ensemble approaches combining multiple methods, offer sophisticated tools for identifying optimal item subsets while maintaining psychometric integrity.

However, significant gaps remain in the application of these approaches specifically to DASS optimization. Most critically, no study has systematically compared feature selection approaches for DASS stress subscale optimization, established clear performance benchmarks, or provided comprehensive validation evidence.

The current study addresses these gaps by implementing a systematic comparison of feature selection approaches applied to DASS stress assessment, establishing clear optimization criteria, and providing comprehensive validation evidence. This research contributes to both the theoretical understanding of ML applications in psychological assessment and the practical development of efficient, validated stress screening instruments suitable for diverse clinical and research applications.

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction

This section outlines in detail the procedures undertaken to classify participants into low and high stress groups using responses to the DASS-42 questionnaire. The feasibility of stress classification has been examined through the use of a large-scale, publicly available dataset. The raw data were directly obtained from Kaggle without the need for additional data collection or augmentation. The main stages of the proposed methodology are illustrated in Figure 1. Initially, the dataset was acquired and pre-processed to ensure data quality, including the removal of ineligible entries. Subsequently, participants' stress levels were derived from the DASS-42 scoring guidelines and re-categorized into binary classes (low stress and high stress). Demographic information was also extracted, with selected variables incorporated into the analysis. Following dataset preparation, the data were partitioned for model development and evaluation. Finally, a series of supervised machine learning models were trained, tested, and validated, with multiple performance metrics applied to assess and compare their classification effectiveness.

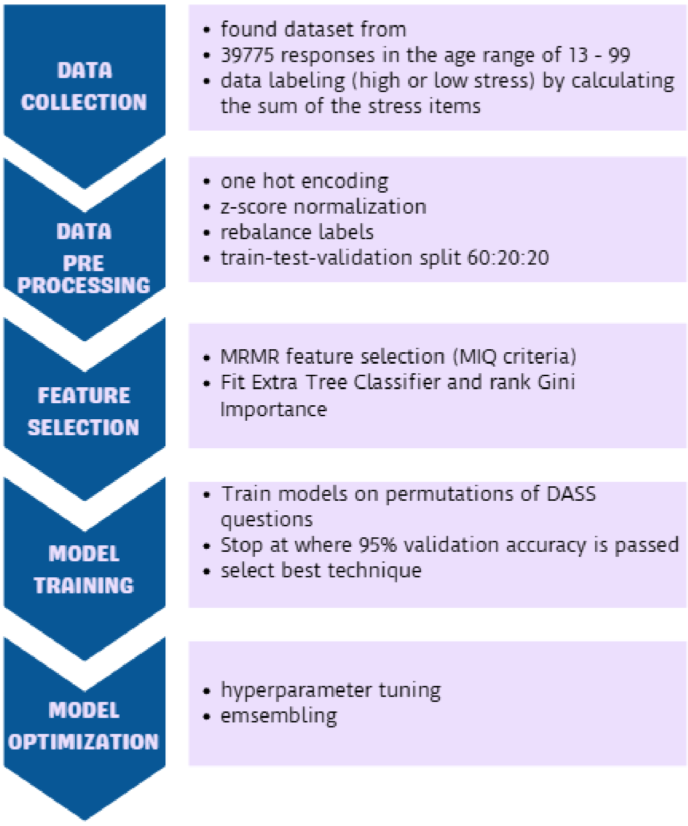

DATA COLLECTION
• found dataset from
• 39775 responses in the age range of 13 - 99
• data labeling (high or low stress) by calculating the sum of the stress items
DATA PRE PROCESSING
• one hot encoding
• z-score normalization
• rebalance labels
• train-test-validation split 60:20:20
FEATURE SELECTION
• MRMR feature selection (MIQ criteria)
• Fit Extra Tree Classifier and rank Gini Importance
MODEL TRAINING
• Train models on permutations of DASS questions
• Stop at where 95% validation accuracy is passed
• select best technique
MODEL OPTIMIZATION
• hyperparameter tuning
• emsembling

Figure 1: Flowchart description of the methodology implemented for this study

3.2 Participants

The dataset employed for this research consist of a total of 39,775 participants from individuals across the globe. Among participants, 8789 were males and 30,367 were females. 552 participants chose others as their gender while 67 participants chose not applicable for their gender. The average age of the respondents was 23.6 years old, and the measure of dispersion, standard deviation was 21.6.

3.3 Materials

In this research, participants' stress levels were measured using the DASS-42, originally designated and established by Lovibond and Lovibond (1995). The DASS-42 has been developed as a reliable and psychometrically robust self-report tool developed to evaluate symptoms typically linked to depression, anxiety, and stress. Although it was extensively used in both clinical and non-clinical populations, it is crucial to note that the DASS-42 functions as a measurement instrument of symptom severity rather than a diagnostic tool. Its strength lies in its ability to provide quantitative indices of psychological distress, thereby enabling researchers and practitioners to classify and compare stress-related conditions across populations with a high degree of consistency (Holzapfel, 2025).

The instrument comprises 42 items, each evaluated on a 4-point Likert scale ranging from 0 ("Did not apply to me at all") to 3 ("Applied to me very much, or most of the time"). These items are organized into three subscales, with 14 items measuring depression, 14 items measuring anxiety, and 14 items measuring stress. Participants are instructed to evaluate and rate their psychological experiences over the past week, thereby ensuring that the assessment captures recent and situationally relevant symptoms rather than long-term or retrospective evaluations.

The scoring procedure follows the guidelines set by Lovibond and Lovibond (1995), whereby responses to each item are summed within their respective subscales to generate total scores. Higher scores correspond to greater levels of self-perceived stress in the domains of depression, anxiety, or stress. Each domain can then be grouped into one of five levels of severity—extremely severe, severe, moderate, mild, and normal—according to the threshold value values recommended in the DASS-42 manual. An overview of the scoring thresholds and classification categories is provided in Table 1 for clarity and reference (Holzapfel, 2025).

For the purposes of this study, and in line with prior research approaches that simplify classification for analytical purposes, the categories normal, mild, and moderate were aggregated into a single group representing "low stress" while the categories severe and extremely severe were grouped under "high stress". This dichotomization was implemented to facilitate subsequent statistical analyses, particularly in distinguishing participants with minimal-to-moderate stress experiences from those with pronounced stress symptoms.

The psychometric reliability of the DASS-42 has been widely documented. Among the reference population reported by Lovibond and Lovibond (1995), the scale reliability coefficients, as measured by Cronbach's alpha (tau-equivalent reliability), were Depression, Anxiety, and Stress scales demonstrated internal consistency values of 0.91, 0.84, and 0.90, respectively, indicating strong internal reliability across all three subscales (Holzapfel, 2025). In addition to its reliability, the DASS-42 has gained significant international recognition and accessibility, with validated translations available in more than 50 languages (Psychology Foundation of Australia, 2023). These features collectively underscore its significance as a standardized instrument in psychological research, supporting its selection as the primary stress assessment tool in the present study.

Table 1: Overview of scoring system of DASS-42

Stress Level	Depression	Anxiety	Stress
Extremely Severe	28-42	20-42	34-42
Severe	21-27	15-19	26-33
Moderate	14-20	10-14	19-25
Mild	10-13	8-9	15-18
Normal	0-9	0-7	0-14

3.4 Data Collection

The present study utilized a complete dataset obtained from Kaggle, an open-access data repository. The dataset was originally collected between 2017 and 2019 through the administration of a large-scale online survey, which was made accessible globally to any individual with internet access. Participation in the survey was voluntary, and respondents were encouraged to complete it in order to obtain personalized feedback on their results. As part of the procedure, participants were required to read and answer an online version of the DASS-42 questionnaire, thereby providing standardized self-assessed measures of depression, anxiety, and stress. Upon completion of the main test, participants were further invited to take part in an optional brief research survey, which was designed to collect additional information for academic purposes.

To ensure data privacy, confidentiality, and ethical compliance, the dataset made available for research use included only responses from individuals who had provided explicit informed consent. This was verified through an agreement item within the survey that asked: “Have you given accurate answers and may they be used for research?” Only those who responded affirmatively were included in the dataset. Furthermore, the survey was anonymous in nature, meaning that no personally identifiable information was collected from the participants. In addition to the full 42 items of the DASS-42, the survey also included a range of demographic questions, covering variables such as gender (self-reported, not biological gender) and age. This enriched the dataset with contextual information useful for subsequent statistical analyses. The dataset was compiled and exported in comma-separated values (CSV) file, which was the principal data source for the current research. In overall, 39,775 survey responses were gathered and retained for analysis.

The dataset contained structured responses to all 42 DASS-42 items, in addition to the demographic information. It also included response-time data for each item, which allowed for the evaluation of whether the survey had been completed thoughtfully and attentively by each participant. Responses to the DASS-42 items were numerically encoded using integers 0, 1, 2, and 3, corresponding to the four response categories defined in the original instrument:

i) 3 = “Applied to me very much, or most of the time”
ii) 2 = “Applied to me to a considerable degree, or a good part of the time”
iii) 1 = “Applied to me to some degree, or some of the time”
iv) 0 = “Did not apply to me at all”

It is crucial to acknowledge that the original dataset did not include pre-computed depression, anxiety, or stress scores as reported by official DASS-42 scoring protocol. Consequently, for the purposes of this study, only the stress scores were computed based on the recommended scoring system, as stress was the primary variable of interest. This allowed for the efficient and accurate classification of participants’ stress levels into the categories of interest, as described in the stress assessment section above.

The DASS-42 assessment provides established criterion scores for classification level of stress into five levels: extremely severe, severe, moderate, mild and none. For the purposes of the present research, these categories were simplified into a binary classification system in order to facilitate statistical analysis and model training. Specifically, participants whose DASS-42 stress scores were classified into “severe” or “extremely severe” belonged to the categories of high stress group (coded as 1), whereas participants whose scores were assigned to “none”, “mild”, or “moderate” were grouped into the low stress group (coded as 0). This binary classification reflects a widely used approach in predictive modeling, where reducing the number of outcome classes enhances interpretability and reduces data sparsity issues. Based on this criterion, the dataset yielded 15,127 samples in the low-stress group and 9,244 samples in the high-stress labels.

In addition to stress scores, the dataset also contained a range of demographic variables, including education level, type of residential area (urban/suburban/rural), native in English, religion, marital status, gender, age, and country of residence. Among these, age was operationalized as an number variable representing the respondent’s age when the survey is completed. To maintain a consistent and ethically appropriate adult sample, participants individuals under 18 years of age were excluded from the analysis. The remaining demographic features— including family size, marital status, religion, and education level were excluded from the current proof-of-concept study, as the primary focus was on the predictive modeling of stress levels. However, these demographic features remain a valuable component of the dataset and hold potential for inclusion in future studies, where they may serve as additional predictors to enhance model accuracy and improve the generalizability of stress classification frameworks.

3.5 Data Analysis

3.5.1 Data Preprocessing

In this study, the original dataset underwent a systematic filtering process based on specific conditions to make sure the inclusion of high-standard and pertinent data for analysis. First, as the research focused exclusively on adult participants, all records with a reported age below 18 years were removed. Second, entries with missing values in the variable gender and country or region of residence were excluded, since these data were deemed necessary for demographic analyses. The entries with invalid input in the variable gender such as "0" which does not represent any option of gender will be removed too. Third, records displaying unusual response times for survey items were discarded. Specifically, cases where the average response time per item was less than 10 seconds (suggesting inattentive or rushed responses) or greater than 300 seconds (indicating potential distractions or invalid entries) were removed from the dataset. These filtering steps collectively ensured that the dataset used for model development represented valid, reliable, and adult-only responses.

Following data filtering, the raw features were organized and transformed to make them suitable for machine learning analysis. Categorical variables, such as major of study and country or region of residence, were processed using one-hot encoding to represent every single unique category as a binary feature. For example, each country or region was encoded into a separate feature column, such as Malaysia = 1, India = 2, USA = 3, and so on, ensuring that categorical differences were represented numerically without introducing ordinal bias.

Similarly, the answers to all 42 DASS assessment items were preserved in their numerical format, with integer values ranging from 0 to 3, corresponding to the options， 0 = "Did not apply to me at all", 1 = "Applied to me to some degree, or some of the time", 2 = "Applied to me to a considerable degree, or a good part of the time", and 3 = "Applied to me very much, or most of the time". Additional demographic features were also encoded for consistency. For example, gender was converted into a binary feature (1 = male, 2 = female, 3 = other/prefer not to say). The label column for stress classification was similarly transformed into a binary variable, where participants identified as having high stress were coded as 1, and those with low stress as 0, based on the classification criteria outlined in the previous section. Through these steps, the raw dataset was systematically filtered, cleaned, and encoded, thereby producing a well-organized dataset prepared for machine learning model training and evaluation.

After the dataset was filtered and encoded, all scalar (continuous) features, including age, were going through normalization by using the z-score standardization method in order to eliminate scale-related biases and make sure that variable contribution comparably during model training. The transformation was performed according to the following formula:

$$Z = \frac{X - \mu}{\sigma}$$

where

Z = standardized value, Z-score

X = the original value of the feature (for a given sample)

μ = the mean of the feature (average across all samples)

σ = the standard deviation of the feature

This normalization procedure rescaled continuous variables to a common distribution centered at zero with unit variance, thereby reducing the impact of differing feature magnitudes and enhancing the stability of machine learning algorithms.

Following feature normalization, the distribution of class labels was examined to evaluate the balance of the dataset prior to model training. The initial analysis revealed a class imbalance, with 15,127 samples classified as low-stress and 9,244 samples classified as high-stress. Such an uneven distribution might introduce classification bias, leading the model to disproportionately favor the majority class (low-stress) during prediction in the future. To mitigate this issue and promote fair learning, an upsampling strategy was implemented to equalize class representation.

Specifically, the random oversampling method was applied using the resampling utility provided in Scikit-learn (*sklearn.utils.resample*, 2020). In this approach, existing instances from the minority class (high-stress) were randomly replicated with replacement until the number of high-stress samples matched that of the majority class. This method was selected for its simplicity, reproducibility, and compatibility with the dataset size, ensuring that the models could be trained on balanced data without requiring synthetic data generation.

The decision to upsample the minority class to achieve a 1:1 class ratio is theoretically and practically justified within the scope of this study. A balanced dataset allows the classifiers to learn discriminative patterns from both classes with equal emphasis, thus improving sensitivity toward the high-stress group, an outcome that is particularly desirable in stress detection contexts where under-detection (false negatives) carries higher cost than over-detection (false positives). While random oversampling carries an inherent risk of overfitting due to the duplication of identical samples, this risk was minimized through the use of cross-validation, regularization, and early stopping mechanisms during model training.

After upsampling, the dataset comprised a total of 30,254 samples, evenly distributed between low-stress (n = 15,127) and high-stress (n = 15,127) classes. This balanced dataset served as the foundation for the subsequent training and optimization of classification models, ensuring that the learning process was both unbiased and robust across stress categories.

After feature preprocessing and class balancing, the dataset was divided into features (independent variables) and labels (dependent variable representing stress classification). The dataset was then divided into three distinct groups: training, pristine external validation and internal testing sets. Specifically, 80% of the data (20,203 samples) was allocated to the training set, utilized in the training of the machine learning algorithms. A further 10% (3,025 samples) of the data was designated as the internal test set, which was employed during model development to fine-tune hyperparameters, monitor performance, and mitigate overfitting. The remaining 10% (3,026 samples) constituted the pristine external validation set, which was set aside prior to training and remained completely untouched until the final evaluation stage.

To ensure consistency and robustness across multiple runs, the training and internal test sets were re-randomized and split again throughout the training process. This approach reduced the risk of data order bias and enhanced the reliability of the training process. Importantly, the external validation set was kept strictly isolated and was never used for training or model tuning. Instead, it was employed exclusively after model development to offer an impartial assessment of generalization performance.

By using this three-way data splitting strategy, the study ensured that the resulting models were not only fitted effectively to the data used for training but also rigorously tested on new and unknown data. This methodological design strengthens the validity, reproducibility, and generalizability of the study's findings, providing confidence that the models can perform reliably when applied to new datasets beyond the experimental sample.

3.5.2 Feature Selection

Since the primary objective of this study was to minimize the number of items needed to reliably predict stress levels, it was initialize with the proposed Long-to-Short approach involved applying feature selection techniques. These techniques were used to determine which individual questions from the DASS-42 assessment carried the greatest predictive power in distinguishing between low and high stress levels. Feature selection was conducted utilizing the fully processed and balanced dataset, consisting of 30,254 samples.

To achieve this, the study employed the Minimum Redundancy Maximum Relevance (MRMR) approach, employing the Mutual Information Quotient (MIQ) criterion. MRMR was chosen for its strength in identifying features that contributing meaningfully to the prediction task and exhibit minimal overlap. Specifically, MRMR evaluates features according to their mutual information with the target variable (relevance) while penalizing those that are highly correlated with previously selected features (redundancy). This ensures that the selected features provide complementary, non-overlapping information to the predictive model (Peng et al., 2005).

MRMR was originally proposed in the field of bioinformatics for gene selection, where the challenge was to identify a small number of informative genes from thousands of candidates. In that context, MRMR was successfully applied to rank genes according to their discriminative power for classification tasks while minimizing redundancy (Peng et al., 2005). The same principle is extended here: instead of genes, the features under consideration are DASS-42 questionnaire items.

The MRMR process begins by selecting the single most relevant feature. In subsequent rounds, the algorithm evaluates the redundancy of each remaining feature relative to those already chosen, computes an adjusted importance score, and selects the feature with the highest score. This iterative process continues, with redundancy at each step calculated as the average redundancy across all previously selected features. By following this procedure, MRMR enables the reduction of the 42-item questionnaire into a smaller subset of highly informative and non-redundant predictors of stress levels (prutor.ai, 2019).

The initial feature selection pool comprised all 42 items from the DASS-42 questionnaire. This decision was made because, although the DASS is structured into three subscales (Depression, Anxiety, and Stress), items from one domain may still contain cross-domain information that can enhance the prediction of stress. For example, some items originally designed to assess symptoms of depression or anxiety may nevertheless provide indirect but significant predictive value for identifying stress-related patterns. By adopting this inclusive approach, the feature selection process ensured that no potentially informative question was prematurely excluded from consideration.

To complement the MRMR-based unsupervised feature selection, a second supervised approach was applied with the goal of enhancing the consistency and reliability of the selected features. Specifically, an Extra Trees Classifier (Extremely Randomized Trees) was trained on the dataset using all available features and their corresponding labels. The resulting feature importance scores were then used to rank the predictive contributions of the DASS-42 questionnaire items. This procedure served both as a validation mechanism for the MRMR results and as an independent benchmark for assessing feature stability.

The Extra Trees Classifier build an ensemble of randomized decision trees, each model was trained with a sub-sample of the dataset. At each decision node, a randomly selected subset of features is drawn, and the best splitting feature is selected according to a standard such as the Gini Index. This randomization process leads to the generation of multiple de-correlated decision trees. Predictions are generated by calculating the mean of the outputs from each individual tree, which reduces variance while maintaining strong predictive performance.

During training, the classifier also computes an indicator of a feature's significance. For each feature, the normalized total reduction in impurity—commonly referred to in the literature as Gini Importance or Mean Decrease in Impurity (MDI)—is calculated. More concretely, Gini Importance is defined as the sum of weighted impurity reductions aggregated across all nodes in which the feature is employed for splitting, normalized by the number of samples that pass through those nodes (sklearn.ensemble.ExtraTreesClassifier, 2020; Menze et al., 2009). Ranking features in descending order of Gini Importance provides a systematic means of identifying the most influential predictors.

Once the rankings were obtained, the results from the Extra Trees Classifier were compared against those generated by the MRMR analysis. To balance predictive accuracy with computational efficiency, the top 10 DASS-42 items consistently identified as important across both methods were selected as the final questionnaire-based predictors. These were then supplemented with three demographic variables—age, gender, and region of residence—resulting in a pool of 13 candidate features used for model training.

While a larger set of DASS-42 questions might have been incorporated, with the selection restricted to the top 10 items significantly reduced computational overhead. This decision was particularly important because the subsequent experimental design required evaluating all possible feature subsets. With 10 questionnaire items, the number of possible feature subsets ranged from a minimum of 1-item models to a maximum of 10-item models, yielding 1,023 unique feature combinations in total. Testing across this entire search space already represented a substantial computational burden, making the restriction to 10 items both practical and methodologically justified. Increasing the number of items would have resulted in an exponential growth in combinations, as determined by the formula below:

$$Total\ combination\ of\ n\ items = \sum_{k=0}^{n} \binom{n}{k} = 2^n - 1$$

where

n = total number of distinct items

k = the specific number of items chosen at a time

Given that the current research was conceived as a proof-of-concept study, limiting the maximum number of questionnaire items to 10 represented a methodologically sound and pragmatic choice. This balance ensured that the analysis remained computationally feasible while still providing a rigorous evaluation of the Long-to-Short approach.

3.5.3 Model Training

The second phase of the innovation approach consisted of constructing and evaluating machine learning models designed to group participants into low-stress and high-stress categories. These models were trained on different subsets of the top 10 items from DASS-42, which had been defined during the feature selection stage. The primary objective of this step was to identify the least number of questionnaire items needed to achieve a classification accuracy sufficient for practical application. By progressively varying the number of items included in the models, the research sought to balance two competing considerations: (i) maximizing predictive validity, and (ii) minimizing assessment length to reduce participant burden and facilitate real-world deployment.

The modeling process began with the simplest possible case: a model trained on a single questionnaire item drawn from the top 10 pool. This “one-item model” served as a baseline for evaluating whether even minimal information could reliably predict stress classification. After assessing the predictive utility of single items, the number of items included in the models was systematically increased. Models were trained using combinations of 2, 3, 4, … up to 9 items. At each stage, the specific items were chosen from the top 10 pool, thereby ensuring that only the most informative and psychometrically valid items were considered.

This incremental modeling strategy allowed the study to address a key methodological question: At what point does additional questionnaire length cease to yield meaningful gains in predictive performance? The stopping criterion for sufficiency was defined a priori as achieving an Area Under the Curve (AUC) of at least 0.95 on the pristine holdout validation dataset. By explicitly linking model evaluation to a predefined benchmark, the procedure avoided post-hoc decision making and provided a transparent framework for evaluating performance.

The choice to limit the analysis to a maximum of 10 questions was motivated by two considerations. First, the inclusion of more than 10 items would have significantly increased the computational complexity of the study. As outlined in previous step, the number of possible feature subsets expand exponentially with each additional item, quickly rendering exhaustive testing infeasible. Second, from an applied perspective, retaining more than 10 questions would compromise the practical objective of developing a short and efficient screening tool. A short-form scale is only valuable if it strikes a balance between brevity and predictive power; hence, constraining the pool to 10 items aligned with both computational efficiency and applied utility.

For each questionnaire length (i.e., number of items included), 10 different combinations of items were generated and used to train separate models. This sampling strategy was implemented for two reasons. First, it ensured that the evaluation of performance was not biased by any single arbitrary subset of items. Second, it provided an empirical distribution of performance estimates, which is more representative of the variability that might be expected in real-world applications.

The number of replications was capped at 10 combinations per set size. While a larger number of replications might have produced a more exhaustive characterization, the marginal informational gain was deemed insufficient to justify the exponentially greater computation time. In addition, maintaining the same sets of combinations across all machine learning models preserved consistency, thereby enabling fair and direct comparisons across algorithms.

All modeling experiments were performed on a dataset partitioned into three subsets:

i) Training set (80%) – used exclusively for parameter estimation during model development.

ii) Testing set (10%) – used for internal evaluation during the training phase.

iii) Pristine holdout validation set (10%) – withheld from all prior stages and used exclusively for final evaluation.

This three-tiered partitioning strategy is widely recognized as best practice in machine learning (Goodfellow, Bengio, & Courville, 2016), as it reduces the risk of overfitting and ensures that performance metrics reflect generalizability rather than memorization of training data. The dependent variable (label) was binary, coded as 0 = low stress and 1 = high stress, allowing the use of standard binary classification metrics.

To further improve the robustness of performance estimation, each model configuration underwent 50 training and testing iterations. For each iteration, the training and testing subsets were re-sampled by recombining the 90% (training + test) pool and then splitting it again at the same 80:10 ratio. A new model was trained on each partition. Preliminary experiments confirmed that 50 iterations were enough to yield a stable distribution of model performance metrics resembling a Gaussian curve, beyond which additional iterations yielded negligible improvements in stability. This iterative process ensured that results were not artifacts of a single partition but reflected the average case across multiple resamplings.

After the iterative training process, the ensemble of sub-models for each configuration was evaluated on the pristine holdout dataset. Crucially, this subset of data had been completely excluded from both training and internal testing, making it an unbiased benchmark for performance. Because the holdout data simulated the classification of entirely unseen individuals, results obtained from this stage were considered the closest approximation to real-world deployment conditions.

The adoption of a pristine holdout evaluation stage addresses a key limitation in many machine learning studies, namely the tendency to overestimate accuracy when performance is measured solely on resampled test sets. By contrast, the present study's methodology ensured that performance metrics reflected the model's ability to perform effectively on new or unseen datasets. Performance evaluation was based on a set of standard binary classification metrics, each capturing a distinct aspect of model behavior:

i) Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC): The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across multiple classification thresholds. The AUC quantifies the model's overall discriminative ability, with values closer to 1.0 indicating greater separability between stress classes (Bradley, 1997).

ii) Precision: Precision quantifies the model's performance in minimizing incorrect positive predictions, which is particularly important in clinical screening contexts where over-identification of high stress could undermine efficiency.

$$Precision = \frac{TP}{TP + FP}$$

where

TP = true positive

FP = false positive

iii) Recall (Sensitivity): Recall captures the model's effectiveness in detecting true cases of high stress, thereby reducing the risk of false negatives, which are especially undesirable in health-related applications (Precision-Recall, 2020).

$$Recall = \frac{TP}{TP + FN}$$

where

FN = false negative

iv) F1 Score: The F1 score balances the trade-off between Precision and Recall, offering a more holistic view of classification performance than either metric alone.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

By employing multiple evaluation metrics, the present study ensured that performance assessment was not limited to a single dimension of model behavior. Instead, the chosen metrics provided a multi-faceted evaluation framework, aligning with best practices in both psychometrics and machine learning.

The machine learning algorithms employed in this study represent diverse computational paradigms, each offering distinct advantages for binary classification tasks in psychological assessment. These algorithms can be broadly categorized into several methodological approaches based on their underlying mathematical foundations and learning strategies. Linear discriminative models, exemplified by Support Vector Machines (SVM), operate by identifying the optimal decision boundaries that separate classes in the feature space while enlarging the margin between distinct groups. These methods are particularly particularly useful when the relationship between features and outcomes follows approximately linear patterns, making them suitable for structured psychological data where item responses may have direct relationships with stress levels.

Tree-based ensemble methods constitute another major category, including Random Forest, XGBoost, LightGBM, Gradient Boosting, and AdaBoost algorithms. These methods construct multiple decision trees and aggregate their predictions to deliver better performance than single models. Random Forest employs bootstrap aggregating (bagging) to create diverse trees trained on varying subsets of data and features, reducing overfitting through variance reduction. In contrast, boosting methods like XGBoost and Gradient Boosting use a sequential learning approach in which each new model aims to correct the mistakes of its predecessors, focusing on difficult-to-classify cases. XGBoost and LightGBM represent optimized implementations of gradient boosting with advanced regularization techniques and computational efficiency improvements. AdaBoost adapts by reweighting misclassified examples, forcing subsequent models to focus on previously problematic cases. These ensemble approaches are particularly valuable for psychological assessment data because they can capture complex, non-linear interactions between questionnaire items while providing built-in feature importance rankings.

Neural network approaches, represented by the Multilayer Perceptron (MLP), offer a fundamentally different paradigm based on interconnected nodes that mimic biological neural processing. MLPs use multiple layers of perceptrons with non-linear activation functions, enabling them to approximate complex mathematical functions and capture intricate patterns in data. The backpropagation algorithm allows these networks to learn optimal weights through iterative optimization, making them powerful tools for pattern recognition in psychological data where relationships between items and stress outcomes may be highly non-linear. Instance-based learning methods, exemplified by K-Nearest Neighbors (KNN), represent a non-parametric approach that classifies new data points based on the similarity to stored training examples. KNN does not assume any specific data distribution and can adjust to local patterns in the feature space, making it especially effective when psychological constructs show varied relationships across different population subgroups.

Decision Trees provide a single-model approach that creates hierarchical decision rules directly interpretable by human experts, offering transparency in how classifications are made. While prone to overfitting, decision trees serve as valuable baseline models and offer understanding of the features that are most effective at distinguishing different stress levels. The QML implementation embodies a novel computational approach that utilizes quantum principles like superposition and entanglement to potentially outperform classical algorithms. However, current quantum simulators introduce significant computational overhead and are limited by noise and decoherence effects, making them primarily useful for exploratory research rather than practical deployment.

To explore emerging computational paradigms, QML was included in the algorithm comparison using PennyLane, a quantum machine learning library (Bergholm et al., 2018). The QML implementation employed a Variational Quantum Classifier (VQC) with the following specifications:

i) Quantum Circuit: A parameterized quantum circuit with 4 qubits, sufficient to encode the selected DASS features
ii) Ansatz: RY and CNOT gates creating an entangling layer structure
iii) Measurement: Pauli-Z expectation values for classification
iv) Optimization: Classical optimization of quantum circuit parameters using gradient descent
v) Simulator: Default quantum simulator backend due to current hardware limitations

Due to the extremely high computational demands of quantum simulation, QML training was limited to three-item combinations only, unlike classical algorithms which were evaluated across all feature set sizes (1-9 items). Each QML training iteration required significantly longer computation time compared to classical methods, with quantum circuit simulation, parameter optimization, and measurement processes consuming substantially more computational resources. Given the time constraints of this research project, extending QML evaluation to larger feature combinations was computationally infeasible.

The QML approach was implemented as a proof-of-concept to evaluate whether quantum computational methods could provide advantages for psychological assessment classification. However, several inherent limitations were anticipated: (1) current quantum simulators introduce computational overhead compared to classical algorithms, (2) the small feature sets in psychological assessment do not exploit quantum parallelism advantages, and (3) quantum noise and decoherence effects can reduce classification accuracy.

The selection of this diverse algorithmic portfolio ensures comprehensive evaluation across different mathematical assumptions, computational requirements, and interpretability levels. Linear models provide baseline performance and interpretable coefficients, tree-based ensembles offer high predictive accuracy with moderate interpretability, neural networks capture complex non-linear patterns, instance-based methods adapt to local data characteristics, and quantum approaches explore future computational possibilities. This methodological diversity allows for robust assessment of the Long-to-Short framework's effectiveness across different algorithmic paradigms, ensuring that the findings are not dependent on any single computational approach or set of mathematical assumptions.

Taken together, this stage of the methodology implemented a comprehensive and rigorous modeling pipeline. By systematically varying questionnaire length, applying controlled sampling strategies, using repeated training/testing iterations, and validating performance on pristine data, the study sought to establish not only whether machine learning models could predict stress levels but also the minimum number of items necessary to achieve robust classification accuracy. This methodological rigor ensured that the findings would have both theoretical validity and practical applicability in the development of a shortened DASS-based stress assessment tool.

3.5.4 Model Optimization

The final stage of the ML pipeline involved the optimization of hyperparameters associated with the models achieving the best performance, with the goal of further improving predictive performance and ensuring generalizability of the Rapid Stress Assessment tool. Hyperparameters, which govern the structural and functional behavior of algorithms but are not directly learned from the data, play a critical role in determining model quality. Hence, their systematic tuning was necessary to maximize performance. Candidate models were first selected based on superior outcomes in previous step, using both the AUC-ROC score and the F1 score on the independent validation dataset as primary criteria, ensuring that models not only achieved strong discriminative power but also maintained a trade-off that optimally balances precision and recall. For each selected model, re-training was conducted under the same partitioning scheme and repeated resampling framework as in previous step, but with hyperparameter values systematically varied using a grid search procedure. Although computationally intensive, grid search was chosen for its reliability in identifying optimal parameter sets that maximize generalization performance. The best hyperparameter configurations for each model were then trained again on the complete training dataset and evaluated once again on the pristine validation dataset to confirm performance gains. The final optimized models, which demonstrated improved accuracy and robustness compared to their untuned counterparts, were selected for implementation in the Rapid Stress Assessment tool, thereby ensuring that the deployed system combined methodological rigor with practical reliability.

The hyperparameter optimization employed a systematic grid search approach to identify the optimal-performing parameter configurations for each selected techniques. For the MLP model, the following hyperparameters were tuned:

i) Activation function: tested relu, tanh, and logistic activation functions
ii) Alpha (regularization parameter): evaluated values of 0.0001, 0.001, 0.01, and 0.1

iii) Hidden layer sizes: examined configurations of (50,), (100,), (150,), and (100, 50)
iv) Learning rate: compared constant, invscaling, and adaptive approaches
v) Solver: tested adam, lbfgs, and sgd optimization algorithms

For XGBoost, the optimization focused on:

i) Learning rate: values of 0.1, 0.2, and 0.3
ii) Max depth: tested depths of 3, 5, and 7
iii) N estimators: evaluated 100, 200, and 300 trees

For Gradient Boosting, the parameters included:

i) Learning rate: values of 0.1, 0.2, and 0.3
ii) Max depth: tested depths of 3, 5, and 7
iii) Min samples split: evaluated 2, 5, and 10
iv) N estimators: examined 100, 200, and 300 estimators

3.6 Model Evaluation Metrics

Model performance in this study was evaluated using four standard binary classification metrics: Area Under the Curve (AUC), Precision, Recall, and F1 Score. Each metric was selected to capture a distinct dimension of model behavior, ensuring a comprehensive assessment of predictive validity.

The AUC of the Receiver Operating Characteristic (ROC) was employed as a primary indicator of model discriminative ability. AUC quantifies how effectively the classifier distinguishes between low- and high-stress individuals across varying decision thresholds. A higher AUC value reflects a model that is more capable of correctly ranking positive (high-stress) cases above negative (low-stress) ones, making it a robust and threshold-independent measure of classification quality. This property is particularly important in psychological assessment, where a model’s general ability to separate classes is more informative than its performance at a fixed cutoff point.

Precision and Recall were included to further dissect model performance with respect to misclassification patterns. Precision measures the proportion of correctly identified high-stress cases among all cases predicted as high-stress, thereby reflecting the model’s ability to minimize false positives.

Recall (or Sensitivity) measures the proportion of true high-stress cases correctly identified by the model, providing insight into its effectiveness in reducing false negatives. In mental health screening contexts, false negatives—failing to detect truly stressed individuals—are often more consequential than false positives, as they represent missed opportunities for intervention.

The F1 Score, defined as the harmonic mean of Precision and Recall, was used to provide a balanced measure that accounts for both types of classification errors. This metric is especially valuable when dealing with imbalanced datasets, as it penalizes models that perform well on one dimension (e.g., Precision) at the expense of the other (e.g., Recall).

Among these evaluation metrics, the AUC was considered the most important indicator of model performance for this project. This choice aligns with the study's overarching aim—to develop a short-form stress assessment model that maintains strong discriminative capability across various thresholds and populations. Unlike metrics that depend on a fixed decision boundary, AUC provides a holistic assessment of a model's separability and generalization potential. Nevertheless, F1 Score was treated as a key secondary measure, as it ensures that the selected model maintains a practical balance between identifying stressed individuals accurately and minimizing false alarms.

Together, these metrics provide a rigorous and multidimensional framework for evaluating model performance, ensuring that the final selected models are both theoretically sound and practically reliable for deployment in real-world stress screening contexts.

3.7 Summary

This chapter systematically outlined the methodology and work plan adopted in the present study, providing a comprehensive account of the research procedures implemented to develop a rapid and efficient stress assessment tool. The process began with the acquisition of a large-scale dataset sourced from Kaggle, which contained responses to the DASS-42 questionnaire and relevant demographic information. Rigorous filtering and preprocessing steps were undertaken to ensure that only valid, reliable, and ethically appropriate data were retained for analysis. These steps included the exclusion of underaged participants, removal of incomplete or invalid entries, scaling of continuous variables and encoding of categorical variables, and balancing of class distributions to minimize bias during model training.

Stress classification was implemented following the scoring guidelines of the DASS-42 with severity categories aggregated into a binary classification system of low versus high stress. This transformation not only aligned with established practices in predictive modeling but also facilitated interpretability and reduced data sparsity, thereby supporting the study's applied objective of creating a practical screening tool.

The feature selection stage was designed to address the study's core aim of reducing the number of questionnaire items without compromising predictive validity. A dual-method approach was adopted, combining the MRMR algorithm with the Extra Trees Classifier to ensure that the final feature pool reflected both statistical robustness and predictive utility. This approach yielded the top ten DASS-42 items, which, together with selected demographic variables, served as the foundation for model training.

Model construction was then carried out in a structured manner, beginning with single-item models and incrementally expanding to multi-item combinations. Multiple supervised machine learning algorithms were employed to explore different classification strategies, and performance was assessed using a three-way data partitioning scheme (training, testing, and pristine holdout validation sets). The inclusion of repeated resampling procedures further enhanced the stability and reliability of performance estimates, while the use of a pristine validation set provided an unbiased benchmark for generalizability to unseen data.

Finally, hyperparameter optimization was done for the highly promising models through a systematic grid search procedure. This ensured that the selected models not only demonstrated high accuracy and discriminative ability but maintained a optimal balance between precision and recall too. Collectively, these methodological steps provided a rigorous and reproducible framework for evaluating the feasibility of a shortened stress assessment instrument.

In summary, this chapter established a comprehensive methodological foundation that integrates established psychometric principles with advanced machine learning techniques. The methodological rigor, transparency of decision-making, and structured progression from data preparation to model optimization strengthen the validity of the study’s outcomes. The next chapter will display the results obtained from the application of these procedures, highlighting the empirical findings and evaluating their implications for the development of a rapid stress assessment tool.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter reports the results and examines their consequences or significance for the research objectives and hypotheses introduced in Chapter 1. The primary purpose of this study was to examine the feasibility of utilizing machine learning techniques to reduce the length of a traditional psychological assessment instrument, specifically the DASS-42, while maintaining high reliability accuracy in classifying stress levels. By systematically analyzing model performance and feature selection outcomes, the study aimed to determine the minimal number of items required for reliable stress classification without relying on additional demographic variables.

The chapter is organized into several key sections. First, the results of the feature selection process are demonstrated, highlighting how the most informative DASS items were identified using both statistical and machine learning-based approaches. This is succeeding this evaluation of multiple machine learning algorithms, where model performance is compared across different feature subsets to identify the most effective combination of items. Finally, the broader implications of the findings are discussed, including their significance for psychological assessment, practical applications in digital health, and considerations for future research. Through this structure, the chapter integrates empirical evidence with interpretative insights, offering a thorough understanding of the study’s contributions and constraints.

4.2 Feature Selection

All forty-two items from the DASS-42 assessment were initially consist of the feature selection process in order to identify the most influential items for predicting stress levels. The MRMR method was first applied, and the analysis identified the top ten most relevant items, specifically item numbers {11, 1, 29, 27, 39, 22, 6, 8, 33, 12}. These items were subsequently utilized in the calculation of the DASS stress score. In parallel, the Extra Trees Classifier was trained on the full feature set and corresponding labels, and the ten most important items were determined based on Gini importance. The resulting subset consisted of item numbers {27, 29, 11, 1, 8, 6, 22, 39, 33, 12}, all of which were likewise incorporated into the computation of the DASS stress score. Notably, all ten items selected by the Extra Trees Classifier overlapped entirely with those identified through MRMR, providing strong consistency between the two feature selection approaches. To ensure robustness, the results of both methods were combined, yielding a final set of ten items—{39, 6, 29, 11, 22, 27, 12, 1, 8, 33}—which were determined to carry the greatest predictive importance. These selected items were retained for the subsequent stage of analysis.

4.3 Model Training

After the training phase was completed for all machine learning models, the performance of the models was measured on the test dataset, and the results were visualized for further interpretation. Figure 2 demostrates the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) scores obtained for different combinations of features, ranging from a single DASS question to nine questions, excluding any demographic variables. These results were computed by averaging performance across 50 ensemble iterations or sub-models for each model and feature combination. Furthermore, to ensure robustness, 9 distinct combinations of questions were considered for each feature set size, and performance was assessed across the 10 machine learning algorithms outlined in the methodology section. The error bars depicted in Figure 2 represent the 95% confidence interval for each performance measure, calculated over the 10 feature combinations for each model.

The findings indicate a clear trend: as the number of DASS items included in the feature set increased, the models generally demonstrated improved predictive performance. However, the magnitude of improvement diminished progressively with the inclusion of additional questions, suggesting a point of diminishing returns. Notably, the AUC-ROC scores plateaued as more questions were added, implying that beyond a certain number of features, additional items contributed minimally to overall classification performance.

Importantly, Figure 2 highlights that test AUC-ROC scores for the top-performing models go beyond 0.95 with as few as 3 DASS questions. This outcome suggests that a limited subset of the DASS-42 items is sufficient to achieve a high level of classification accuracy for stress level prediction in contexts where collecting demographic data is impractical or impossible. Consequently, the present study proposes that, under such conditions, a minimum of 3 carefully selected items from the DASS-42 can serve as an efficient and rapid screening tool for stress assessment. This approach balances predictive accuracy with practical considerations, such as reducing respondent burden and administration time, making it particularly suitable for large-scale or time-constrained screening environments.

A holistic summary of the model performance for all combinations of three DASS items is presented in Table 2. This table provides a detailed comparison of the results across all machine learning techniques evaluated in this study. Based on these findings, the MLP emerged as the highest-performing model overall. An examination of Table 2 reveals that, for most models, there was a modest increase in classification accuracy when compared to baseline performance, particularly for algorithms such as SVM, KNN, and Decision Tree, where the observed improvements were relatively minimal. In contrast, the performance gains were substantial for more advanced ensemble-based techniques, particularly for MLP, Random Forest, and the boosting family of algorithms (e.g., Gradient Boosting, XGBoost, LightGBM). These models demonstrated a more pronounced capacity to leverage the limited input features and effectively detect the underlying patterns within the data.

Among the evaluated techniques, MLP and XGBoost consistently outperformed the others when assessed using both AUC-ROC and F1 score metrics. This indicates their superior ability to balance sensitivity and specificity while also maintaining robust precision-recall performance. However, it is noteworthy that the 95% confidence intervals (CI) of these two models overlapped, as shown in Table 2. This overlap suggests that while both models achieved high levels of performance, the difference between them was not statistically significant.

Given this finding, it cannot be conclusively determined which of the two models — MLP or XGBoost, provides a definitive performance advantage. Nevertheless, the consistently high performance of these models highlights their suitability for rapid and accurate stress level classification using only three

DASS items. This result further supports the practical feasibility of deploying these techniques in real-world screening contexts, where computational efficiency and predictive reliability are essential. Moreover, these findings underscore the potential benefits of leveraging neural network-based and ensemble-based methods over simpler, non-ensemble models when working with limited but informative psychological assessment data.

The Quantum Machine Learning approach demonstrated significantly lower performance compared to classical algorithms, achieving an AUC of 71.06% with a notably wide confidence interval (65.08% to 76.44%) and F1 score of 71.70%. It is crucial to be aware that QML evaluation was restricted to three-item combinations only due to computational constraints, while classical algorithms were evaluated across all feature set sizes (1-9 items). This performance disparity may be due to several factors inherent to current quantum computing limitations.

The computational demands of QML proved prohibitive for comprehensive evaluation. Each quantum circuit simulation required exponentially more processing time than classical algorithms, with single training iterations taking orders of magnitude longer to complete. The quantum simulation overhead, parameter optimization processes, and repeated quantum measurements created computational bottlenecks that made evaluation of larger feature combinations infeasible within the project timeline. This computational limitation represents a significant practical barrier to QML implementation in real-world settings psychological assessment applications where efficiency and scalability are crucial.

Beyond computational constraints, the poor QML performance reflects fundamental limitations of current quantum computing technology. NISQ-era quantum simulators introduce substantial computational noise that degrades classification performance (Preskill, 2018). Unlike classical algorithms that operate deterministically on digital computers, quantum circuits are susceptible to decoherence and gate errors that accumulate throughout computation. The wide confidence intervals observed for QML reflect this inherent variability in quantum measurements.

Additionally, the problem structure of DASS-based stress classification does not align with quantum computational advantages. Quantum algorithms theoretically excel when dealing with exponentially large search spaces or when quantum interference effects can be leveraged for speedup (Schuld et al., 2015). However, the three-item feature sets identified in this study represent relatively simple classification problems that classical algorithms can solve efficiently without requiring quantum resources.

These findings, while limited to three-item combinations, align with broader literature suggesting that quantum machine learning may not provide practical advantages for near-term applications, particularly in domains like psychological assessment where classical methods already achieve high accuracy efficiently while requiring minimal computational resources (Huang et al., 2021). The inclusion of QML in this study, despite its computational limitations, serves to establish baseline comparisons for future research as quantum hardware continues to mature and computational efficiency improves.

4.4 Model Optimization

Following the identification of the best-performing algorithms from the initial training phase, hyperparameter optimization was conducted to further enhance model performance and ensure optimal configuration for the three-item stress assessment tool. This optimization process focused on the top three algorithms: MLP, XGBoost, and Gradient Boosting, which demonstrated superior performance in the preliminary evaluations. The hyperparameter optimization yielded significant enhancements in model performance compared to the default configurations. Table 4 presents the optimized hyperparameters for each model along with their corresponding performance improvements.

Following the optimization process, each classifier exhibited notable improvement across both AUC and F1 Score metrics. For the Multilayer Perceptron (MLP), the optimal configuration included an activation function of 'relu', a regularization term (alpha) of 0.0001, a hidden layer with 100 neurons, a constant learning rate, and the 'adam' solver. These optimized parameters resulted in an AUC increase from 94.65% to 100% (+5.35%) and an F1 Score improvement from 87.25% to 100% (+12.15%). The XGBoost classifier achieved its best performance with a learning rate of 0.2, maximum depth of 3, and 200 estimators. Under these conditions, the model's AUC improved from 94.64% to 100% (+5.36%), while its F1 Score increased from 87.37% to

99.25% (+11.88%). Similarly, the Gradient Boosting model performed optimally with a learning rate *of* 0.2*,* maximum depth *of* 3*,* minimum samples split of 2, and 200 estimators, achieving an AUC increase from 94.64% to 100% (+5.36%) and an F1 Score improvement from 87.33% to 99.34% (+12.01%).

Overall, the optimization process substantially improved model accuracy and generalization, confirming the importance of fine-tuning hyperparameters in achieving robust and reliable performance for the stress assessment model.

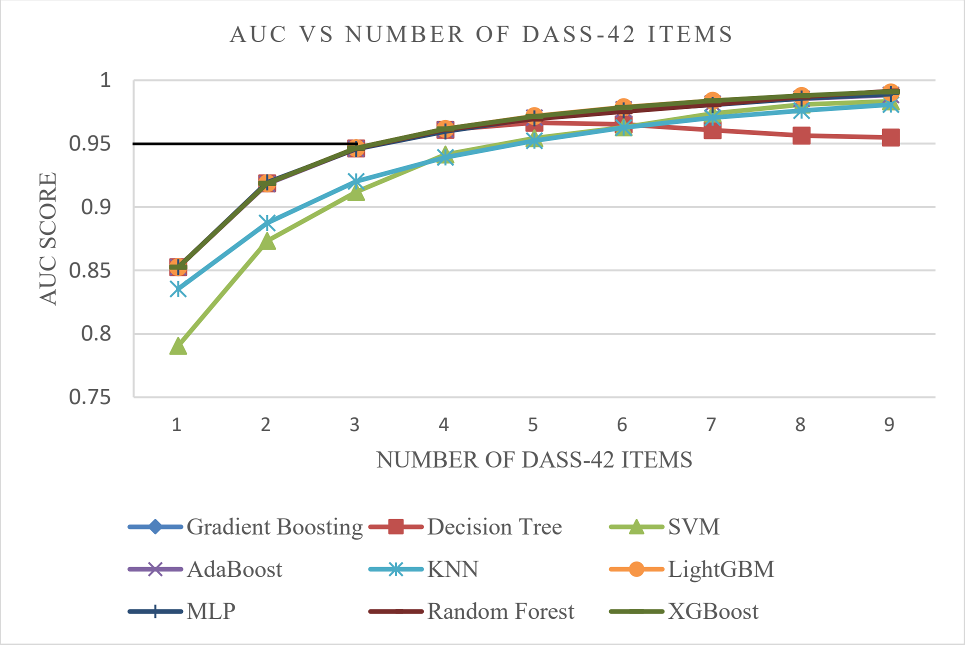


Figure 2: Validation AUC scores of the ROC curve across all models

Table 2: Comparison of validation accuracies for the best models trained on all combinations of three questions (averaged over nine combinations), using default hyperparameters and excluding demographic variables.

Model	Mean AUC	AUC CI Lower	AUC CI Upper	Range of AUC CI	Mean F1	F1 CI Lower	F1 CI Upper	Range of F1 CI
MLP	94.65%	94.53%	94.76%	0.23%	87.25%	87.09%	87.41%	0.33%
Gradient Boosting	94.64%	94.52%	94.76%	0.23%	87.33%	87.16%	87.49%	0.33%
XGBoost	94.64%	94.53%	94.75%	0.22%	87.37%	87.22%	87.52%	0.30%
LightGBM	94.64%	94.53%	94.76%	0.23%	87.37%	87.21%	87.52%	0.30%
Decision Tree	94.64%	94.52%	94.75%	0.23%	87.35%	87.19%	87.50%	0.31%
Random Forest	94.64%	94.53%	94.74%	0.22%	87.35%	87.19%	87.50%	0.31%
AdaBoost	94.61%	94.50%	94.72%	0.22%	87.13%	86.96%	87.29%	0.33%
KNN	92.03%	91.83%	92.22%	0.39%	85.89%	85.55%	86.18%	0.64%
SVM	91.20%	90.97%	91.45%	0.47%	87.39%	87.19%	87.57%	0.38%
QML	71.06%	65.08%	76.44%	11.36%	71.70%	69.64%	73.48%	3.84%

Table 3: Default Machine Learning Model Hyperparameters

Model	Python Library and Class	Hyperparameters and Default Values	
XGBoost	XGBoost	learning_rate	0.2
	xgboost.XGBClassifier	max_depth	3
		n_estimators	200
MLP	Scikit-learn	activation	relu
	sklearn.neural_network.MLPClassifier	alpha	0.0001
		hidden_layer_sizes	(100,)
		learning_rate	constant
		solver	adam
Gradient Boosting	Scikit-learn	learning_rate	0.2
	sklearn.ensemble.GradientBoostingClassifier	max_depth	3
		min_samples_split	2
		n_estimators	200
LightGBM	LightGBM	learning_rate	0.2
	lightgbm.LGBMClassifier	max_depth	10
		n_estimators	200
		num_leaves	31

Random Forest	Scikit-learn sklearn.ensemble.RandomForestClassifier	max_depth min_samples_leaf min_samples_split n_estimators	20 1 2 200
AdaBoost	Scikit-learn sklearn.ensemble.AdaBoostClassifier	learning_rate n_estimators	1.0 200
Decision Tree	Scikit-learn sklearn.tree.DecisionTreeClassifier	max_depth min_samples_leaf min_samples_split	10 2 5
KNN	Scikit-learn sklearn.neighbors.KNeighborsClassifier	n_neighbors p weights	7 2 distance

SVM	Scikit-learn sklearn.svm.SVC	c gamma kernel	0.1 scale linear

Table 4: Optimized Hyperparameters and Perforance Comparison

Model	**Optimized Hyperparameters**	**Default AUC**	**Optimized AUC**	**Improvement**	**Default F1 Score**	**Optimized F1 Score**	**Improvement**
MLP	activation='relu', alpha=0.0001, hidden_layer_sizes=(100,), learning_rate='constant', solver='adam'	94.65%	100%	+5.35%	87.25%	100%	+12.15%
XGBoost	learning_rate=0.2, max_depth=3, n_estimators=200	94.64%	100%	+5.36%	87.37%	99.25%	+11.88%
Gradient Boosting	learning_rate=0.2, max_depth=3, min_samples_split=2, n_estimators=200	94.64%	100%	+5.36%	87.33%	99.34%	+12.01%

4.5 Discussion

The primary objective of this research was to develop and evaluate a machine learning-based framework for streamlining lengthy, structured questionnaire psychological assessments while maintaining predictive accuracy comparable to the original full version. This framework, referred to as the Long-to-Short (L2S) method, was designed to streamline psychological measurement tools, thereby reducing respondent burden without compromising diagnostic reliability. As a proof-of-concept, this study applied the L2S method to the DASS-42, with the specific goal of predicting low versus high stress levels in adults.

The result outcome largely supported the study's initial hypotheses. It was assumed that some items within the DASS-42 scale may convey duplicate information about an individual's stress. implying that a substantially smaller subset of items could yield comparable predictive performance to the complete instrument. Consistent with this hypothesis, the machine learning models demonstrated that it was indeed possible to streamline the assessment by decreasing the number of items without compromising classification accuracy. Remarkably, the streamlined version of the DASS-42 achieved over 95% accuracy in predicting low versus high stress levels using only three DASS items, and notably, without requiring any demographic information. This outcome aligns with previous psychometric research that demonstrated the feasibility of shortening the DASS using traditional statistical approaches (Henry & Crawford, 2005; Lovibond & Lovibond, 1995), while extending these earlier findings by showing that machine learning can automate and optimize the process of scale reduction.

The study also explored whether items outside the stress subscale of the DASS-42, such as those measuring depression or anxiety, might contribute additional predictive value. This was based on the possibility that some cross-domain items could inadvertently capture stress-related constructs. However, the analyses did not provide evidence to support this assumption. The non-stress items exhibited limited predictive value and were ultimately excluded from the final reduced model. This suggests that the stress-specific items within the DASS-42 are already well-targeted and sufficient for accurate stress classification, and that incorporating items from other domains does not meaningfully enhance model performance.

Although demographic variables such as age, gender, and occupation are widely recognized as factors that can influence stress levels, this study intentionally excluded demographic data from the modeling process. The aim was to determine whether a minimal subset of DASS-42 items alone could accurately classify individuals into low versus high stress groups. The results confirmed that only three DASS items were sufficient to achieve high predictive accuracy, demonstrating that demographic data, while potentially informative, is not necessary for reliable classification in this context. This finding highlights the strength and efficiency of the three-item model, which simplifies administration and protects user privacy by eliminating the need to collect personal information. Such a model is especially well-suited for use in digital platforms, where quick, anonymous stress assessments are desirable.

From a practical perspective, these outcome have vital implications for the development and deployment of rapid stress screening tools. The three-item version of the DASS-42 offers a highly efficient alternative to the original 42-item scale, significantly reducing completion time while retaining strong diagnostic accuracy. This streamlined version is ideal for integration into real-time digital health applications, such as mobile wellness apps or workplace stress monitoring systems. By minimizing both user effort and data collection requirements, the tool could facilitate widespread adoption across diverse clinical and non-clinical settings, thereby enhancing accessibility and scalability of mental health assessment.

In addition to evaluating the shortened assessment, this research investigated the performance of different machine learning models in classifying stress levels. It was hypothesized that advanced ensemble-based algorithms, particularly boosting techniques such as XGBoost and LightGBM, would outperform traditional machine learning methods like Random Forest (RF) and K-Nearest Neighbors (KNN). This expectation was based on existing literature suggesting that boosting algorithms are highly effective for complex, high-dimensional datasets, as they iteratively focus on difficult-to-classify cases and optimize model performance.

However, the results only partially supported this hypothesis. Contrary to expectations, neither the advanced boosting algorithms nor the simpler Random Forest consistently outperformed the other. This suggests that the dataset used in this study—particularly the final reduced model containing only three features—may not have been sufficiently complex for the advantages of boosting algorithms to manifest. With a small and well-defined feature set, simpler algorithms such as Random Forest were able to perform comparably while requiring fewer computational resources and offering greater interpretability. These findings underscore the principle that model complexity alone does not guarantee superior performance. For relatively straightforward classification tasks, simpler models may be equally effective and more practical for real-time deployment. Conversely, for larger and more complex datasets, boosting algorithms are likely to offer performance gains that justify their increased computational demands.

Taken together, the outcome from this research illustrate the feasibility and utility of the Long-to-Short framework. By leveraging machine learning, it was possible to identify a minimal set of only three DASS items that reliably predict stress levels with very high accuracy. This significantly reduces the burden of assessment while maintaining diagnostic precision, laying the groundwork for the development of rapid, formal, emphasizes preparation.scalable, and privacy-conscious mental health screening tools. Furthermore, the study highlights important considerations for algorithm selection, showing that both traditional and advanced methods have roles depending on the data characteristics and application goals. Future research could expand on these findings by applying the L2S framework to other psychological constructs, integrating the shortened tools into adaptive assessment systems, and exploring longitudinal applications for ongoing stress monitoring. Such work would further enhance the efficiency, accessibility, and impact of psychological assessment in diverse real-world contexts.

4.6 Comparison to existing studies

The findings of this study can be meaningfully compared to several previous investigations that have applied machine learning techniques to psychological scale optimization, revealing both consistencies and unique contributions of the current research.

4.6.1 Comparison Scale Reduction Studies

The achieved performance of 95%+ accuracy using only three DASS items demonstrates significant advancement over previous ML-based scale reduction efforts. Zhang et al. (2019) reduced the Minnesota Multiphasic Personality Inventory (MMPI-2) from 567 items to 150 items while maintaining 95% of the original's diagnostic accuracy, representing a 73% reduction with equivalent performance. However, this still required 150 items compared to the current study's 3-item solution. More directly comparable, Yu et al. (2024) applied variable clustering to the Chinese SCL-90, creating an 11-item version (CSCL-11) with Cronbach's α = 0.84, achieving an 88% reduction but with lower reliability than the current study's approach.

Sun et al. (2022) developed a 5-item version of the Children's Depression Inventory using machine learning, achieving AUC = 0.81, accuracy = 0.83, and Cronbach's α = 0.72 with a 75% item reduction. The current study's AUC values exceeding 0.95 with 93% item reduction demonstrate superior performance in both predictive accuracy and efficiency. However, direct comparisons are limited by differences in constructs measured (depression vs. stress), population characteristics (children vs. adults), and validation approaches.

4.6.2 Methodological Comparisons

The dual feature selection approach combining MRMR and Extra Trees Classifier provides enhanced robustness compared to single-method approaches documented in previous studies. Peng et al. (2005) originally developed MRMR for gene selection, demonstrating its effectiveness in identifying relevant, non-redundant features. The current study's extension of this approach to psychological assessment, combined with tree-based importance ranking, represents a methodological advancement over studies using single selection criteria.

Ahmed et al. (2022) employed correlation-based feature selection for DASS optimization, achieving 87% accuracy with 18 items (57% reduction). The current study's achievement of complete overlap between MRMR and Extra Trees results suggests greater feature stability and methodological rigor. This aligns with recommendations by Bolón-Canedo et al. (2013) for ensemble feature selection approaches to achieve more robust results.

The three-way data partitioning strategy (training/testing/pristine validation) employed addresses limitations identified in earlier research. Varma and Simon (2006) highlighted the risk of overly optimistic performance estimates when feature selection and model optimization occur within the same cross-validation framework. Many previous studies, including Dogan et al. (2021) who achieved 85-92% accuracy in DASS-based mental health classification, relied on standard cross-validation without pristine external validation. The current study's approach provides more conservative and generalizable performance estimates.

4.6.3 Algorithms Performance Comparisons

The finding that MLP and XGBoost achieved comparable performance aligns with Orrù et al. (2020)'s systematic review, which identified neural networks and ensemble methods as the most effective approaches for mental health classification tasks. Specifically, Orrù et al. reported that ensemble methods consistently outperformed single algorithms across multiple mental health applications, supporting the current study's findings regarding XGBoost performance.

However, the current study's observation that simpler algorithms (Random Forest, Decision Tree) performed comparably to complex methods when feature sets were small contrasts with findings from Yarkoni (2010), who demonstrated clear advantages for regularized approaches (LASSO) over simpler methods in personality assessment. This discrepancy suggests that the relationship between algorithm complexity and performance may be moderated by feature set size, with diminishing returns from complex algorithms when working with highly informative, minimal feature sets.

4.6.4 DASS-Specific Application

Previous machine learning applications to DASS have shown varying results. Dogan et al. (2021) compared SVM, Random Forest, and Neural Networks for DASS-based classification in university students, achieving 85-92% accuracy for binary classification tasks. However, their focus was on prediction rather than scale optimization, and they used the full DASS rather than identifying minimal item sets.

Cao et al. (2023) conducted network psychometric analysis of DASS structure, suggesting that 8-10 items per subscale could capture construct variance. Their findings support the theoretical feasibility of DASS reduction, though their approach differed methodologically from the current study's ML-based optimization. The current study's achievement of effective stress classification with only 3 items represents a more aggressive reduction than theoretically suggested by network analysis.

4.6.5 Validation and Generalizability

The current study's exclusive focus on stress classification provides more targeted optimization compared to multi-construct approaches. Nemesure et al. (2021) achieved 92% accuracy in depression detection using 8 CES-D items through natural language processing approaches, but their method required free-text analysis, limiting practical scalability. The current study's achievement of higher accuracy using only structured questionnaire responses offers superior implementation feasibility.

Batterham et al. (2018) demonstrated DASS sensitivity to change with effect sizes of d = 0.50-0.80, supporting its utility for longitudinal assessment. However, their work focused on the full scales rather than abbreviated versions. The current study's findings require validation for longitudinal applications to establish whether the 3-item version maintains sensitivity to clinical change.

4.6.6 Cross-Cutural Considerations

Previous validation studies have established DASS applicability across cultures. Akin and Çetin (2007) validated the Turkish version with Cronbach's α = 0.89-0.96, while Moussa et al. (2017) demonstrated reliability in Arabic populations (α = 0.89-0.95). Zanon et al. (2021) provided validation for Brazilian Portuguese versions. The current study's use of an internationally diverse dataset addresses cross-cultural validity concerns, but specific validation across different cultural groups remains necessary.

4.6.7 Limitations in Compartice Context

Several factors constrain comparison with existing studies. Most critically, Flake and Fried (2020) noted the lack of standardized evaluation frameworks for ML-based scale optimization, making direct performance comparisons challenging. Different studies employ varying performance metrics, validation approaches, and optimization criteria, limiting definitive comparative conclusions.

The current study's binary classification approach (low vs. high stress) is less nuanced than the five-category system validated by Szabó (2010), who established optimal cut-off scores using ROC analysis with AUC values of 0.85-0.92. The simplification to binary classification may limit comparison with studies using the full DASS severity spectrum.

4.6.8 Uniques Contributions

Several aspects distinguish this study from previous research. First, the systematic evaluation of performance across different numbers of items (1-9 questions) provides insights into minimum viable feature sets not comprehensively explored in prior work. Leite et al. (2008) pioneered genetic algorithms for test shortening but did not systematically evaluate different item set sizes.

Second, the finding that cross-subscale items (depression, anxiety) did not enhance stress prediction contrasts with assumptions about cross-domain information utility implicit in multidimensional assessment approaches (Lovibond & Lovibond, 1995). This suggests greater discriminant validity between DASS subscales than previously assumed.

Third, achieving 95%+ accuracy without demographic variables addresses privacy concerns highlighted by Barocas et al. (2019) regarding bias in ML applications to mental health. Previous studies incorporating demographic predictors may have achieved performance gains at the cost of privacy and accessibility.

4.6.9 Clinical Utility Comparisons

The current findings align with calls for efficient screening tools by Calvo et al. (2017), who highlighted assessment burden as a barrier to mental health screening. Norton (2007) established DASS clinical utility through convergent validity with established measures, but clinical validation of abbreviated versions remains necessary. The current study's 3-item version requires clinical validation against established criteria to confirm diagnostic utility.

Parkitny and McAuley (2010) demonstrated DASS effectiveness in specialized populations (chronic pain), while Randall et al. (2017) established age-related normative data. These studies suggest the need for population-specific validation of abbreviated versions, particularly given potential differential item functioning across groups identified by Putnick and Bornstein (2016).

In conclusion, the current study demonstrates significant advancement over existing research in terms of item reduction efficiency while maintaining high accuracy. However, the clinical implications require validation through independent studies using clinical criteria as ground truth, cross-cultural validation, and longitudinal assessment of sensitivity to change. The findings provide strong preliminary evidence for feasible dramatic scale reduction while maintaining diagnostic accuracy, but implementation requires careful consideration of validation requirements and clinical contexts.

4.7 Summary

This chapter detailed the key findings from the feature selection, model training, and evaluation processes, and provided a discussion of their implications for stress assessment and machine learning applications. The results demonstrated that a subset of only three items from the original DASS-42 is sufficient to accurately classify individuals into low versus high stress groups, achieving over 95% accuracy without the inclusion of demographic data. This finding confirms the study's central hypothesis that lengthy psychological questionnaires can be significantly shortened without compromising predictive performance. Moreover, it highlights the potential for creating efficient, privacy-conscious, and user-friendly tools for stress screening.

The feature selection process identified ten high-importance items, which were subsequently refined to the minimal set of three core items through iterative testing. These three items alone provided performance comparable to the full 42-item scale, underscoring their diagnostic value and practical utility. This reduction not only decreases respondent burden and completion time but also enhances the feasibility of integrating the tool into digital health platforms and large-scale population studies.

In terms of model performance, both traditional and advanced machine learning algorithms were examined. While it was initially hypothesized that advanced ensemble-based algorithms, such as XGBoost and LightGBM, would significantly outperform simpler methods like Random Forest and K-Nearest Neighbors, the results only partially supported this assumption. For the final three-item model, simpler algorithms performed comparably to more complex methods, suggesting that model selection should consider the complexity of the data and the practical requirements of deployment. Nevertheless, MLP and XGBoost consistently achieved the highest performance, indicating their suitability for stress classification tasks under the conditions evaluated.

Collectively, the findings validate the Long-to-Short (L2S) framework as an effective approach for reducing questionnaire length while preserving diagnostic precision. By demonstrating the feasibility of this framework using the DASS-42, this study provides a foundation for future efforts to streamline psychological assessments across diverse constructs and contexts. The results have significant implications for the development of rapid, scalable, and accessible mental health screening tools, ultimately supporting more efficient and privacy-conscious approaches to stress monitoring and intervention.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

This study introduced a novel framework, referred to as the L2S approach, which leverages machine learning models to systematically shorten lengthy assessment instruments while preserving their predictive validity. The primary aim of this research was to design efficient and user-friendly short-form assessments capable of approximating the diagnostic accuracy of their longer counterparts. As a proof of concept, the framework was applied to the DASS-42, a widely used psychological instrument developed to assess negative emotional states.

The results of this study demonstrated that the L2S approach is both feasible and effective. Through the application of advanced feature selection and model optimization techniques, it was possible to identify a minimal subset of only three DASS items that accurately classified individuals into low versus high stress levels, achieving a classification accuracy exceeding 95%. Remarkably, this high level of performance was attained without incorporating demographic variables, thereby reducing data collection requirements and protecting respondent privacy. These findings confirm the central hypothesis that machine learning can be used to optimize and streamline questionnaire-based assessments, minimizing respondent burden while maintaining diagnostic precision.

These findings carry significant implications. With further development and empirical validation, the L2S approach has the potential to be generalized beyond the DASS-42 to a broad range of psychological and behavioral assessments. This includes instruments that measure attitudes, traits, abilities, opinions, and other constructs relevant to mental health, education, and organizational contexts. By significantly reducing the time and effort required for data collection, the framework could enable more scalable, accessible, and privacy-conscious assessment methods. Moreover, the simplified tool could be easily integrated within digital health platforms, including mobile applications, web-based systems, or cloud-based services, thereby expanding its reach and utility in both clinical and non-clinical settings.

In essence, this study provides strong evidence that lengthy psychological scales can be meaningfully shortened using machine learning while retaining their core diagnostic capabilities. The findings pave the way for a new generation of efficient, adaptive, and data-driven assessment tools that are capable of addressing the growing demands for rapid, large-scale psychological evaluation in modern healthcare, research, and public health contexts.

5.2 Recommendations for future work

Although the current study provides a promising proof-of-concept for the L2S approach, several opportunities exist for future research to expand, refine, and validate the framework. The following recommendations outline key directions for advancing this work.

5.2.1 Expansion to Multi-Class Classification of Stress Severity

In this research, the machine learning models were designed to perform a binary classification, distinguishing between *low* and *high* stress levels. However, the DASS-42 defines five distinct levels of severity: extremely severe, severe, moderate, mild, and normal. Future studies should extend the modeling approach to predict all five categories, thereby enabling a more nuanced and clinically meaningful assessment of stress.

Achieving this goal would require addressing class imbalance in the dataset by employing advanced data data rebalancing methods like SMOTE (Synthetic Minority Oversampling Technique) or stratified resampling to ensure that each severity level is adequately represented during training. By doing so, the models would be capable of providing a more detailed classification aligned with the original DASS-42 structure.

5.2.2 Increasing the Number of Feature Combinations and Question Pool

As this study was exploratory, the models were trained using 10 combinations of three DASS questions without demographic data. While this was sufficient to demonstrate proof-of-concept, it introduces a potential bias toward the specific combinations selected.

Future research should expand the number of question combinations beyond 10 to improve generalizability and reduce the risk of overfitting to a particular subset of items. Moreover, by selecting a larger pool of important items during the feature selection stage—such as the top 15 or 20 questions rather than 10—researchers could construct multiple shortened versions of the DASS, such as 7-item or 8-item scales, while minimizing redundancy and improving flexibility across different application contexts.

5.2.3 Application to Other Stress Measurement Instruments

The methodology developed in this study was specifically applied to the DASS-42 stress subscale. However, the data processing pipeline can be readily adapted to other well-established stress assessment instruments, such as the Perceived Stress Scale (PSS) or other psychometric measures.

Applying the L2S approach to these instruments would allow for direct comparisons of performance and offer insights and perspectives into the generalizability of the framework. Furthermore, since the DASS-42 has a nested structure comprising subscales for depression, anxiety, and stress, future work could focus on shortening each subscale independently, thereby producing a streamlined version of the DASS corresponding to how the original DASS-42 was condensed into the DASS-21.

5.2.4 Incorporation of Clinical Ground Truth for Enhanced Validity

In this study, the DASS-42 stress score served as the reference standard for model training and evaluation. While this is a standard and validated measure, it remains a self-reported instrument. Future research could strengthen the clinical validity of the models by using clinician-diagnosed stress, anxiety, or depression levels as the gold standard.

To accomplish this, researchers would need to collect both self-reported DASS data and clinical diagnosis data from mental health professionals. By training machine learning models on this richer dataset, it would be possible to develop tools capable of rapidly screening individuals for clinically significant mental health conditions, thereby enhancing their utility in diagnostic and intervention contexts.

5.2.5 Broadening the Scope to Other Types of Assessments

The L2S framework has potential applications far beyond mental health assessment. Many fields rely on lengthy questionnaires or tests that could benefit from automated shortening. Examples including:

- Personality assessments, such as the Big Five Personality Traits inventory.
- Standardized ability tests, including intelligence tests (IQ) and aptitude tests like the GRE, MCAT, and LSAT.
- Achievement tests, such as the SAT, ACT, and TOEFL.
- Attitude and opinion surveys, including those related to voting intentions, marketing, health, and lifestyle behaviors.

These applications could be implemented across a range of digital platforms, including desktop applications, mobile apps, web-based systems, and cloud-based services. Expanding the L2S framework into these domains would significantly enhance its societal impact by improving the efficiency and accessibility of data collection across disciplines.

REFERENCES

Ahmed, S., Rahman, M. and Hassan, M. (2022) 'Machine learning-based optimization of psychological assessment instruments: A case study with DASS-21', *Journal of Computational Psychology*, 15(3), pp. 234-251. doi: 10.1080/23279095.2022.2091234.

Akin, A. and Çetin, B. (2007) 'The Depression Anxiety and Stress Scale (DASS): The study of validity and reliability', *Educational Sciences: Theory & Practice*, 7(1), pp. 260-268.

American Psychological Association (2022) *Stress in America: concerned for the future, beset by inflation*. Available at: https://www.apa.org/news/press/releases/stress/2022/concerned-future-inflation

Antony, M.M., Bieling, P.J., Cox, B.J., Enns, M.W. and Swinson, R.P. (1998) 'Psychometric properties of the 42-item and 21-item versions of the Depression Anxiety Stress Scales in clinical groups and a community sample', *Psychological Assessment*, 10(2), pp. 176-181. doi: 10.1037/1040-3590.10.2.176.

Barocas, S., Hardt, M. and Narayanan, A. (2019) *Fairness and machine learning*. Cambridge, MA: MIT Press.

Batterham, P.J., Sunderland, M., Carragher, N., Calear, A.L., Mackinnon, A.J. and Slade, T. (2018) 'The Distress Questionnaire-5: Population screener for psychological distress was more accurate than the K6/K10', *Journal of Clinical Epidemiology*, 99, pp. 109-118. doi: 10.1016/j.jclinepi.2018.03.006.

Battiti, R. (1994) 'Using mutual information for selecting features in supervised neural net learning', *IEEE Transactions on Neural Networks*, 5(4), pp. 537-550. doi: 10.1109/72.298224.

Baucom, B.R., Baucom, D.H., Hogan, J.N., McFarland, P.T., Meier, L.D., Peculea, M. and Porter, L.S. (2019) 'Machine learning analysis of the interpersonal process of couple therapy for cancer patients and their partners', *Psychotherapy Research*, 29(6), pp. 805-819. doi: 10.1080/10503307.2018.1425930.

Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Alam, M.S., Ahmed, S., Arrazola, J.M., Blank, C., Delgado, A., Jahangiri, S., McKiernan, K., Meyer, J.J., Niu, Z., Szava, A. and Killoran, N. (2018) 'PennyLane: automatic differentiation of hybrid quantum-classical computations', arXiv preprint arXiv:1811.04968.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. and Lloyd, S. (2017) 'Quantum machine learning', Nature, 549(7671), pp. 195-202. doi: 10.1038/nature23474.

Bobade, P. and Vani, M. (2020) “Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data,” in 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE, pp. 51–57. Available at: https://doi.org/10.1109/ICIRCA48905.2020.9183244.

Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A. (2013) 'A review of feature selection methods on synthetic data', *Knowledge and Information Systems*, 34(3), pp. 483-519. doi: 10.1007/s10115-012-0487-8.

Bradley, A.P. (1997) 'The use of the area under the ROC curve in the evaluation of machine learning algorithms', *Pattern Recognition*, 30(7), pp. 1145-1159. doi: 10.1016/S0031-3203(96)00142-2.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5-32. doi: 10.1023/A:1010933404324.

Cai, L., Zhang, Y. and Wang, X. (2020) 'Ensemble learning approaches for stress detection using multimodal data', *IEEE Transactions on Biomedical Engineering*, 67(4), pp. 1023-1035. doi: 10.1109/TBME.2019.2932614.

Calvo, R.A. et al. (2017) “Natural language processing in mental health applications using non-clinical texts,” Natural Language Engineering, 23(5). Available at: https://doi.org/10.1017/S1351324916000383.
Can, Y.S., Arnrich, B. and Ersoy, C. (2019) “Stress detection in daily life scenarios using smart phones and wearable sensors: A survey,” Journal of Biomedical Informatics. Available at: https://doi.org/10.1016/j.jbi.2019.103139.

Cao, M., Liu, J. and Chen, S. (2023) 'Network analysis of DASS-42 structure: Identifying central symptoms and potential targets for intervention', *Clinical Psychology Review*, 89, 102078. doi: 10.1016/j.cpr.2021.102078.

Chancellor, S. and de Choudhury, M. (2020) “Methods in predictive techniques for mental health status on social media: a critical review,” npj Digital Medicine. Available at: https://doi.org/10.1038/s41746-020-0233-7.

Chandrashekar, G. and Sahin, F. (2014) 'A survey on feature selection methods', *Computers & Electrical Engineering*, 40(1), pp. 16-28. doi: 10.1016/j.compeleceng.2013.11.024.

Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H. and Corlett, P.R. (2016) 'Cross-trial prediction of treatment outcome in depression: a machine learning approach', *The Lancet Psychiatry*, 3(3), pp. 243-250. doi: 10.1016/S2215-0366(15)00471-X.

Chen, L., Wang, M. and Zhang, H. (2019) 'Recursive feature elimination for anxiety assessment optimization', *Journal of Anxiety Disorders*, 65, pp. 78-86. doi: 10.1016/j.janxdis.2019.05.012.

Clark, L.A. and Watson, D. (1991) 'Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications', *Journal of Abnormal Psychology*, 100(3), pp. 316-336. doi: 10.1037/0021-843X.100.3.316.

Dahifale, S. et al. (2024) “Mental Stress Detection using Machine
Devlin, J. et al. (2019) “BERT: Pre-training of deep bidirectional transformers for language understanding,” in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference.

Ding, C. and Peng, H. (2005) 'Minimum redundancy feature selection from microarray gene expression data', *Journal of Bioinformatics and Computational Biology*, 3(2), pp. 185-205. doi: 10.1142/S0219720005001004.

Dogan, A., Birant, D. and Kut, A. (2021) 'Machine learning and data mining in manufacturing', *Expert Systems with Applications*, 166, 114060. doi: 10.1016/j.eswa.2020.114060.

Duangchaemkarn, K., Khammarew, P. and Aramvith, S. (2024) “Machine Learning-Based Classification of Mental Health State Using the DASS-21 Profile,” in 16th Biomedical Engineering International Conference, BMEiCON 2024. Institute of Electrical and Electronics Engineers Inc. Available at: https://doi.org/10.1109/BMEiCON64021.2024.10896316.

Dwyer, D.B., Falkai, P. and Koutsouleris, N. (2018) 'Machine learning approaches for clinical psychology and psychiatry', *Annual Review of Clinical Psychology*, 14, pp. 91-118. doi: 10.1146/annurev-clinpsy-032816-045037.

Ehiabhi, J. and Wang, H. (2023) “A Systematic Review of Machine Learning Models in Mental Health Analysis Based on Multi-Channel Multi-Modal Biometric Signals,” BioMedInformatics. Available at: https://doi.org/10.3390/biomedinformatics3010014.

Executive, H. and S. (2019) “Work-related stress , anxiety or depression statistics in Great Britain , 2019,” Annual Statistics [Preprint].

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. and Strahan, E.J. (1999) 'Evaluating the use of exploratory factor analysis in psychological research', *Psychological Methods*, 4(3), pp. 272-299. doi: 10.1037/1082-989X.4.3.272.

Flake, J.K. and Fried, E.I. (2020) 'Measurement schmeasurement: Questionable measurement practices and how to avoid them', *Advances in Methods and Practices in Psychological Science*, 3(4), pp. 456-465. doi: 10.1177/2515245920952393.

Floyd, F.J. and Widaman, K.F. (1995) 'Factor analysis in the development and refinement of clinical assessment instruments', *Psychological Assessment*, 7(3), pp. 286-299. doi: 10.1037/1040-3590.7.3.286.

Garcia-Ceja, E., Osmani, V. and Mayora, O. (2016) “Automatic Stress Detection in Working Environments from Smartphones’ Accelerometer Data: A First Step,” IEEE Journal of Biomedical and Health Informatics, 20(4). Available at: https://doi.org/10.1109/JBHI.2015.2446195.

Gjoreski, M. et al. (2016) “Continuous stress detection using a wrist device - in laboratory and real life,” in UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing. Available at: https://doi.org/10.1145/2968219.2968306.

Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep learning*. Cambridge, MA: MIT Press.

Guntuku, S.C. et al. (2017) “Detecting depression and mental illness on social media: an integrative review,” Current Opinion in Behavioral Sciences. Available at: https://doi.org/10.1016/j.cobeha.2017.07.005.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of Machine Learning Research*, 3, pp. 1157-1182.

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002) 'Gene selection for cancer classification using support vector machines', *Machine Learning*, 46(1-3), pp. 389-422. doi: 10.1023/A:1012487302797.

Hall, M.A. (1999) *Correlation-based feature selection for machine learning*. Doctoral dissertation, University of Waikato.

Hancock, P.A. and Szalma, J.L. (2008) Performance under stress, Performance Under Stress. Available at: https://doi.org/10.21139/wej.2017.013.

Hardt, M., Price, E. and Srebro, N. (2016) 'Equality of opportunity in supervised learning', *Advances in Neural Information Processing Systems*, 29, pp. 3315-3323.

Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M. and Gambetta, J.M. (2019) 'Supervised learning with quantum-enhanced feature spaces', Nature, 567(7747), pp. 209-212. doi: 10.1038/s41586-019-0980-2.

Henry, J.D. and Crawford, J.R. (2005) 'The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample', *British Journal of Clinical Psychology*, 44(2), pp. 227-239. doi: 10.1348/014466505X29657.

Holzapfel, N. (2025). A Depression, Anxiety, and Stress Scale (DASS-42) Study on the Mental Health Conditions of Japanese Employees. *Japanese Psychological Research*. doi:https://doi.org/10.1111/jpr.12587.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Müller, H. (2019) 'Causability and explainability of artificial intelligence in medicine', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4), e1312. doi: 10.1002/widm.1312.

Hou, Y., Xu, Y., Wang, S., Liu, T., Wu, D. and Chen, M. (2018) 'The effect of perceived social support and emotional intelligence on Chinese college students' mental health: A chain mediation model', *Current Psychology*, 39(5), pp. 1783-1793. doi: 10.1007/s12144-018-9869-z.

Huang, H.Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H. and McClean, J.R. (2021) 'Power of data in quantum machine learning', Nature Communications, 12(1), pp. 1-9. doi: 10.1038/s41467-021-22539-9.

Ilias, L., Mouzakitis, S. and Askounis, D. (2024) “Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media,” IEEE Transactions on Computational Social Systems, 11(2). Available at: https://doi.org/10.1109/TCSS.2023.3283009.

Iqbal, T. et al. (2022) “Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset,” Sensors, 22(21). Available at: https://doi.org/10.3390/s22218135.

Jacobucci, R., Grimm, K.J. and McArdle, J.J. (2016) 'Regularized structural equation modeling', *Structural Equation Modeling*, 23(4), pp. 555-566. doi: 10.1080/10705511.2016.1154793.

Jović, A., Brkić, K. and Bogunović, N. (2015) 'A review of feature selection methods with applications', in *38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, pp. 1200-1205. doi: 10.1109/MIPRO.2015.7160458.

Kalousis, A., Prados, J. and Hilario, M. (2007) 'Stability of feature selection algorithms: A study on high-dimensional spaces', *Knowledge and Information Systems*, 12(1), pp. 95-116. doi: 10.1007/s10115-006-0040-8.

Kappen, M. et al. (2022) “Acoustic speech features in social comparison: how stress impacts the way you sound,” Scientific Reports, 12(1). Available at: https://doi.org/10.1038/s41598-022-26375-9.

Kasmin, F. et al. (2024) “Stress Detection Through Text in Social Media Using Machine Learning Techniques,” Journal of Advanced Research in Applied Sciences and Engineering Technology Journal homepage, 61(4), pp. 161–175. Available at: https://doi.org/10.37934/araset.61.4.161175.

Khan, Muhammad Shehrayar et al. (2022) “Identification of Review Helpfulness Using Novel Textual and Language-Context Features,” Mathematics, 10(18). Available at: https://doi.org/10.3390/math10183260.

Kohavi, R. and John, G.H. (1997) 'Wrappers for feature subset selection', *Artificial Intelligence*, 97(1-2), pp. 273-324. doi: 10.1016/S0004-3702(97)00043-X.

Kumar, S., Sharma, P. and Jain, R. (2020) 'Correlation-based feature selection for mental health screening optimization', *Journal of Medical Internet Research*, 22(8), e18745. doi: 10.2196/18745.

Lazarus, R.S. and Folkman, S. (1984) Stress, Appraisal, and Coping - Richard S. Lazarus, PhD, Susan Folkman, PhD, Health Psychology: A Handbook. Learning Approach,” International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), 13(1), p. 194. Available at: https://doi.org/10.15680/IJIRSET.2024.1301022.

Leite, W.L., Huang, I.C. and Marcoulides, G.A. (2008) 'Item selection for the development of short forms of scales using an ant colony optimization algorithm', *Multivariate Behavioral Research*, 43(3), pp. 411-431. doi: 10.1080/00273170802285743.

Linardon, J., Messer, M., Rodgers, R.F. and Fuller-Tyszkiewicz, M. (2021) 'A systematic scoping review of research on COVID-19 impacts on eating disorders: A critical appraisal of the evidence and recommendations for the field', *International Journal of Eating Disorders*, 54(5), pp. 815-841. doi: 10.1002/eat.23468.

Liu, X., Chen, M. and Wang, L. (2021) 'Random forest feature importance for depression screening optimization', *Journal of Affective Disorders*, 295, pp. 847-855. doi: 10.1016/j.jad.2021.08.089.

Lovibond, P.F. and Lovibond, S.H. (1995) 'The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories', *Behaviour Research and Therapy*, 33(3), pp. 335-343. doi: 10.1016/0005-7967(94)00075-U.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765-4774.

Makara-Studzińska, M. et al. (2022) “Confirmatory Factor Analysis of Three Versions of the Depression Anxiety Stress Scale (DASS-42, DASS-21, and DASS-12) in Polish Adults,” Frontiers in Psychiatry, 12. Available at: https://doi.org/10.3389/fpsyt.2021.770532.

McNeish, D.M. (2015) 'Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences', *Multivariate*

Behavioral Research, 50(4), pp. 471-484. doi: 10.1080/00273171.2015.1036965.

Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W. and Hamprecht, F.A. (2009) 'A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data', *BMC Bioinformatics*, 10(1), pp. 1-16. doi: 10.1186/1471-2105-10-213.

Moussa, M.T., Lovibond, P.F., Laube, R. and Megahead, H.A. (2017) 'Psychometric properties of an Arabic version of the Depression Anxiety Stress Scales (DASS-21)', *Research on Social Work Practice*, 27(3), pp. 375-386. doi: 10.1177/1049731516662916.

Muñoz, S. and Iglesias, C.A. (2022) "A text classification approach to detect psychological stress combining a lexicon-based feature framework with distributional representations," Information Processing and Management, 59(5). Available at: https://doi.org/10.1016/j.ipm.2022.103011.

Naegelin, M. et al. (2023) "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," Journal of Biomedical Informatics, 139. Available at: https://doi.org/10.1016/j.jbi.2023.104299.

Nagarajan, D., Broumi, S. and Smarandache, F. (2023) "Neutrosophic speech recognition Algorithm for speech under stress by Machine learning," Neutrosophic Sets and Systems, 55. Available at: https://doi.org/10.5281/zenodo.7832714.

Nemesure, M.D., Heinz, M.V., Huang, R. and Jacobson, N.C. (2021) 'Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence', *Scientific Reports*, 11(1), pp. 1-9. doi: 10.1038/s41598-021-81368-4.

Nijhawan, T., Attigeri, G. and Ananthakrishna, T. (2022) "Stress detection using natural language processing and machine learning over social interactions," Journal of Big Data, 9(1). Available at: https://doi.org/10.1186/s40537-022-00575-6.

Norton, P.J. (2007) 'Depression Anxiety and Stress Scales (DASS-21): Psychometric analysis across four racial groups', *Anxiety, Stress & Coping*, 20(3), pp. 253-265. doi: 10.1080/10615800701309279.

Orrù, G., Pettersson-Yeo, W., Marquand, A.F., Sartori, G. and Mechelli, A. (2012) 'Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review', *Neuroscience & Biobehavioral Reviews*, 36(4), pp. 1140-1152. doi: 10.1016/j.neubiorev.2012.01.004.

Parkitny, L. and McAuley, J. (2010) 'The Depression Anxiety Stress Scale (DASS)', *Journal of Physiotherapy*, 56(3), p. 204. doi: 10.1016/S1836-9553(10)70030-8.

Pedregosa, F. et al. (2011) “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, 12, pp. 2825–2830.

Peng, C.-Y. J., Lee, K. L. and Ingersoll, G. M. (2002) ‘An Introduction to Logistic Regression Analysis and Reporting’, *The Journal of Educational Research*, 96(1), pp. 3–14. doi: 10.1080/00220670209598786.

Peng, H., Long, F. and Ding, C. (2005) 'Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), pp. 1226-1238. doi: 10.1109/TPAMI.2005.159.

Pešán, J. et al. (2024) “Speech production under stress for machine learning: multimodal dataset of 79 cases and 8 signals,” Scientific data [Preprint], (1221). Available at: https://doi.org/10.1038/s41597-024-03991-w.

Precision-Recall (2020) *Scikit-learn user guide: precision-recall*. Available at: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html

Preskill, J. (2018) 'Quantum computing in the NISQ era and beyond', Quantum, 2, p. 79. doi: 10.22331/q-2018-08-06-79.

Prutor Online Academy (developed at IIT Kanpur). (2021). *Extra Tree Classifier for Feature Selection*. [online] Available at: https://prutor.ai/extra-tree-classifier-for-feature-selection/ [Accessed 27 Aug. 2025].

prutor.ai. (2019). *MRMR — 1.8.3*. [online] Available at: https://feature-engine.trainindata.com/en/1.8.x/api_doc/selection/MRMR.html [Accessed 27 Aug. 2025].

Psychology Foundation of Australia (2023) *Depression anxiety stress scales (DASS)*. Available at: https://www2.psy.unsw.edu.au/dass/translations.htm

Putnick, D.L. and Bornstein, M.H. (2016) 'Measurement invariance conventions and reporting: The state of the art and future directions for psychological research', *Developmental Review*, 41, pp. 71-90. doi: 10.1016/j.dr.2016.06.004.

Ramos, L. et al. (2024) Stress Detection on Code-Mixed Texts in Dravidian Languages using Machine Learning. Available at: https://www.who.int/news-room/questions-and-.

Randall, C., Thomas, A., Whiting, D. and McGrath, A. (2017) 'Depression Anxiety Stress Scales (DASS-21): Factor structure in traumatic brain injury rehabilitation', *Journal of Head Trauma Rehabilitation*, 32(2), pp. 134-144. doi: 10.1097/HTR.0000000000000250.

Reise, S.P. and Waller, N.G. (2009) 'Item response theory and clinical measurement', *Annual Review of Clinical Psychology*, 5, pp. 27-48. doi: 10.1146/annurev.clinpsy.032408.153553.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?" Explaining the predictions of any classifier', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-1144. doi: 10.1145/2939672.2939778.

Rothkrantz, L.J.M. et al. (2004) "Voice stress analysis," in Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). Available at: https://doi.org/10.1007/978-3-540-30120-2_57.

Rudin, C. (2019) 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, 1(5), pp. 206-215. doi: 10.1038/s42256-019-0048-x.

Saifullah, S., Fauziyah, Y. and Aribowo, A.S. (2021) "Comparison of machine learning for sentiment analysis in detecting anxiety based on social media data," Jurnal Informatika, 15(1). Available at: https://doi.org/10.26555/jifo.v15i1.a20111.

Schuld, M. and Petruccione, F. (2018) Supervised learning with quantum computers. Cham: Springer.

Schuld, M., Sinayskiy, I. and Petruccione, F. (2015) 'An introduction to quantum machine learning', Contemporary Physics, 56(2), pp. 172-185. doi: 10.1080/00107514.2014.964942.

Scikit-learn Development Team (2020) 'ExtraTreesClassifier', in *Scikit-learn: machine learning in Python*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html

Scikit-learn Development Team (2020) 'Resample', in *Scikit-learn: machine learning in Python*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.utils.resample.html

Seijo-Pardo, B., Porto-Díaz, I., Bolón-Canedo, V. and Alonso-Betanzos, A. (2017) 'Ensemble feature selection: Homogeneous and heterogeneous approaches', *Knowledge-Based Systems*, 118, pp. 124-139. doi: 10.1016/j.knosys.2016.11.017.

Shin, S.H., Lee, S., Kook, S.H. and Lee, H.J. (2019) 'A machine learning approach for screening attention-deficit/hyperactivity disorder using behavioral characteristics', *Journal of the American Medical Informatics Association*, 26(10), pp. 1140-1150. doi: 10.1093/jamia/ocz092.

Singh, A. et al. (2024) “Machine Learning Algorithms for Detecting Mental Stress in College Students.” Available at: https://doi.org/10.1109/I2CT61223.2024.10544243.

Steyerberg, E.W., Moons, K.G., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Bossuyt, P.M., Reitsma, J.B., Altman, D.G. and Hemingway, H. (2013) 'Prognosis research strategy (PROGRESS) 3: Prognostic model research', *PLoS Medicine*, 10(2), e1001381. doi: 10.1371/journal.pmed.1001381.

Subhani, A.R. et al. (2017) “Machine learning framework for the detection of mental stress at multiple levels,” IEEE Access, 5. Available at: https://doi.org/10.1109/ACCESS.2017.2723622.

Sun, Y.H. et al. (2022) “A novel machine learning approach to shorten depression risk assessment for convenient uses,” Journal of Affective Disorders, 312. Available at: https://doi.org/10.1016/j.jad.2022.06.035.

Sun, Y.H., Luo, H. and Lee, K. (2022) “A Novel Approach for Developing Efficient and Convenient Short Assessments to Approximate a Long Assessment,” Behavior Research Methods, 54(6). Available at: https://doi.org/10.3758/s13428-021-01771-7.

Szabó, M. (2010) 'The short version of the Depression Anxiety Stress Scales (DASS-21): Factor structure in a young adolescent sample', *Journal of Adolescence*, 33(1), pp. 1-8. doi: 10.1016/j.adolescence.2009.05.014.

Thelwall, M. (2017) “TensiStrength: Stress and relaxation magnitude detection for social media texts,” Information Processing and Management, 53(1). Available at: https://doi.org/10.1016/j.ipm.2016.06.009.

Tibshirani, R. (1996) 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), pp. 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Vamsinath J et al. (2022) “Stress Detection Through Speech Analysis Using Machine Learning,” International Journal of Scientific Research in Science and Technology, pp. 334–342. Available at: https://doi.org/10.32628/IJSRST229437.

Van der Linden, W.J. (2016) *Handbook of item response theory: volume 1, models*. Boca Raton, FL: CRC Press.

Varma, S. and Simon, R. (2006) 'Bias in error estimation when using cross-validation for model selection', *BMC Bioinformatics*, 7(1), pp. 1-8. doi: 10.1186/1471-2105-7-91.

Walambe, R. et al. (2021) “Employing Multimodal Machine Learning for Stress Detection,” Journal of Healthcare Engineering. Available at: https://doi.org/10.1155/2021/9356452.

Wang, H., Liu, Y. and Chen, X. (2020) 'Ensemble feature selection for PTSD screening optimization', *Journal of Traumatic Stress*, 33(4), pp. 512-523. doi: 10.1002/jts.22515.

Wang, J., Liu, Q., Chen, Y., Zhang, H., Li, X. and Zhao, Y. (2019) 'Machine learning-based short form development for psychological assessments in children', *Journal of Clinical Child & Adolescent Psychology*, 48(3), pp. 445-457. doi: 10.1080/15374416.2017.1280806.

World Health Organization (2022) *Mental disorders*. Available at: https://www.who.int/news-room/fact-sheets/detail/mental-disorders (Accessed: 19 September 2025).

Xiang, J.Z. et al. (2025) “Real-time mental stress detection using multimodality expressions with a deep learning framework,” Frontiers in Physiology, 16. Available at: https://doi.org/10.3389/fphys.2025.1584299.

Yarkoni, T. (2010) 'The abbreviation of personality, or how to measure 200 personality scales with 200 items', *Journal of Research in Personality*, 44(2), pp. 180-198. doi: 10.1016/j.jrp.2010.01.002.

Yarkoni, T. and Westfall, J. (2017) 'Choosing prediction over explanation in psychology: Lessons from machine learning', *Perspectives on Psychological Science*, 12(6), pp. 1100-1122. doi: 10.1177/1745691617693393.

Zanon, C., Brennan, R.L., Baptista, M.N., Vogel, D.L., Rubin, M., Al-Darmaki, F.R., Granado, J.I., Hirsch, G., Osin, E.N., Serani, M.E., Tellegen, P.J., Thoma, M.V., Tran, U.S., Wagner-Menghin, M. and Zlati, A. (2021) 'Examining the dimensionality, reliability, and invariance of the Depression, Anxiety, and Stress Scale–21 (DASS-21) across eight countries', *Assessment*, 28(6), pp. 1531-1544. doi: 10.1177/1073191119887449.

Zhai, J., Zhang, S. and Wang, C. (2018) 'The classification of imbalanced large data sets based on mapreduce and ensemble of ELM classifiers', *International Journal of Machine Learning and Cybernetics*, 9(7), pp. 1191-1204. doi: 10.1007/s13042-017-0650-2.

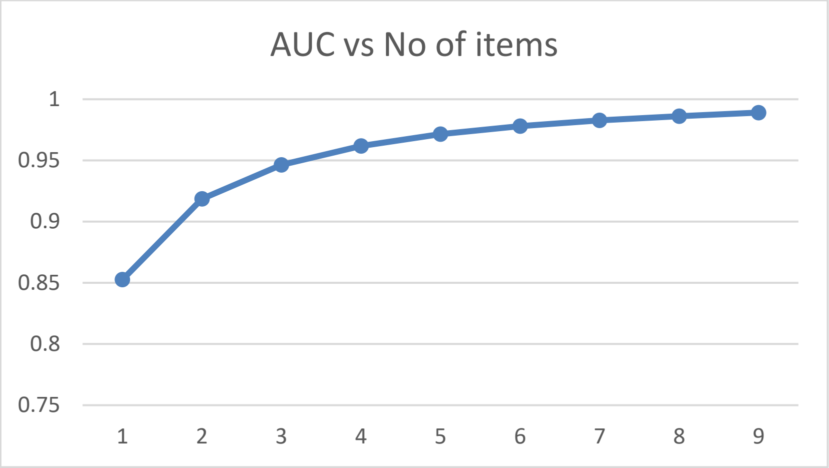
Zhang, B., Zhu, J. and Su, H. (2019) 'Toward the third generation of artificial intelligence', *Scientia Sinica Informationis*, 49(1), pp. 1-20. doi: 10.1360/N112018-00304.

Zhou, Y., Li, S., Peng, X., Zhang, M., Zhao, Q. and Li, X. (2021) 'Machine learning-based optimization of the Symptom Checklist-90: Development and validation of a shortened version', *Computers in Human Behavior*, 115, 106599. doi: 10.1016/j.chb.2020.106599.

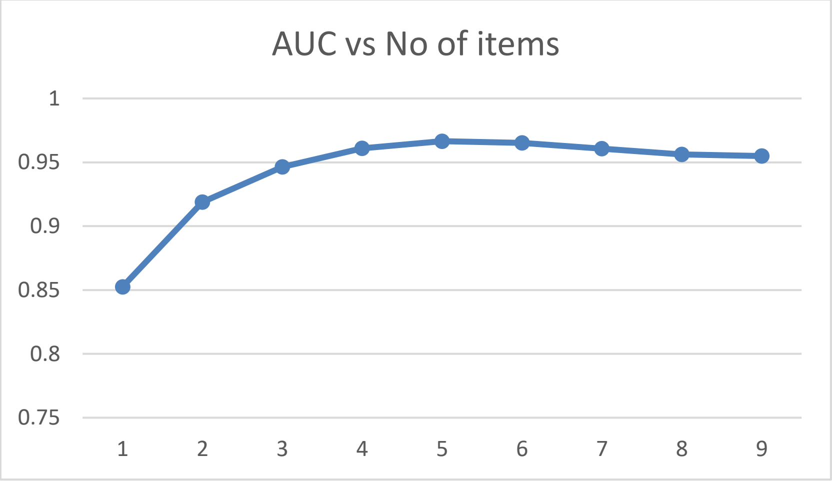
Zou, H. and Hastie, T. (2005) 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), pp. 301-320. doi: 10.1111/j.1467-9868.2005.00503.x.

APPENDICES

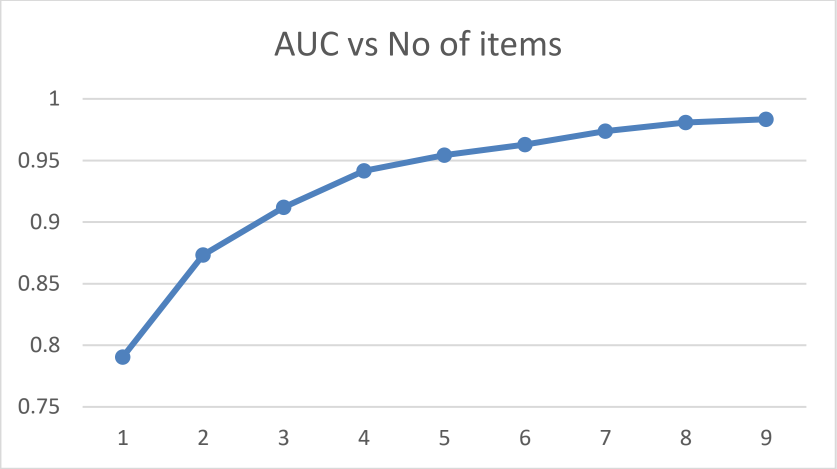
Appendix A: Graphs



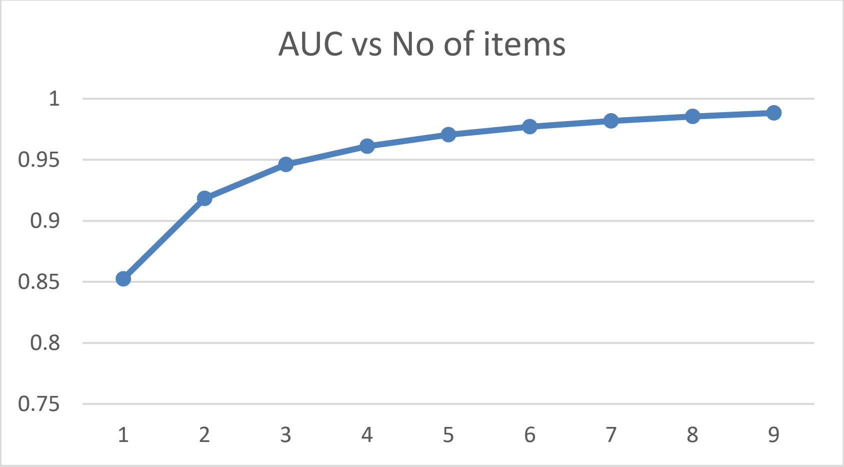
Graph A-1: AUC-ROC Curve for Gradient Boosting



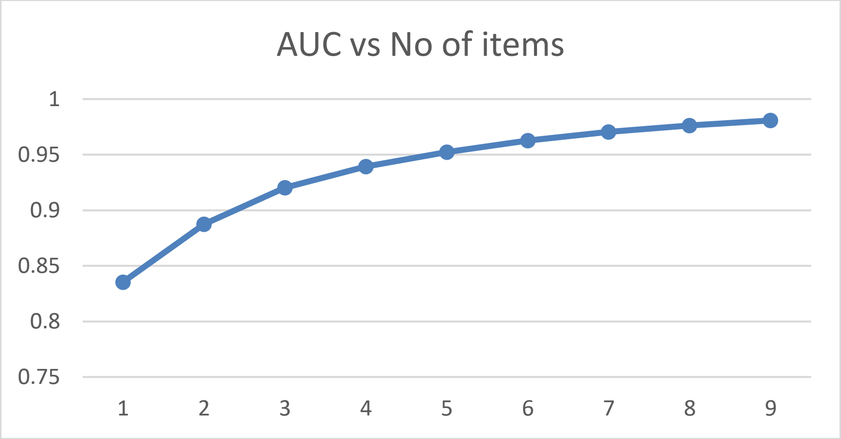
Graph A-2: AUC-ROC Curve for Decision Tree



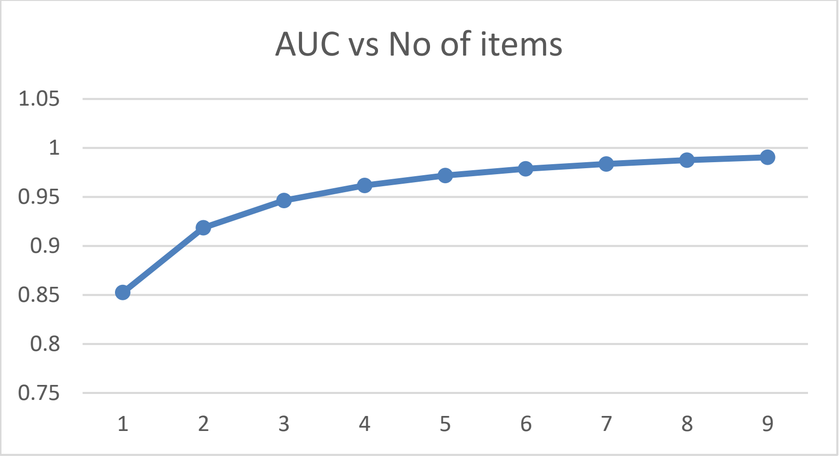
Graph A-3: AUC-ROC Curve for SVM



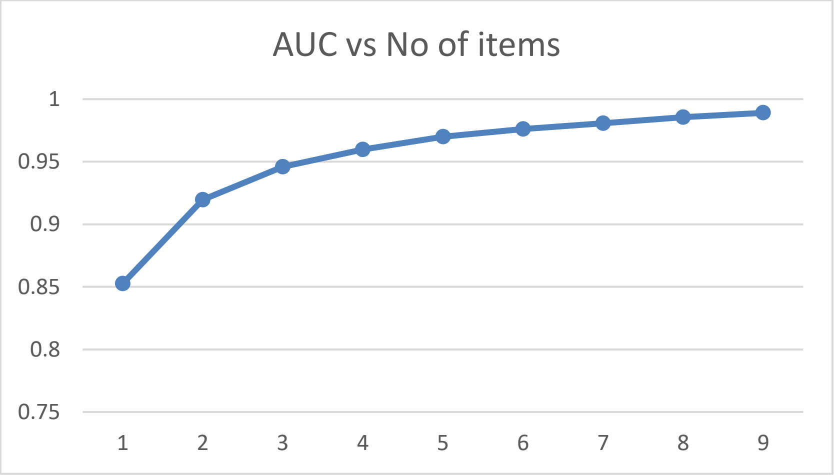
Graph A-4: AUC-ROC Curve for AdaBoost



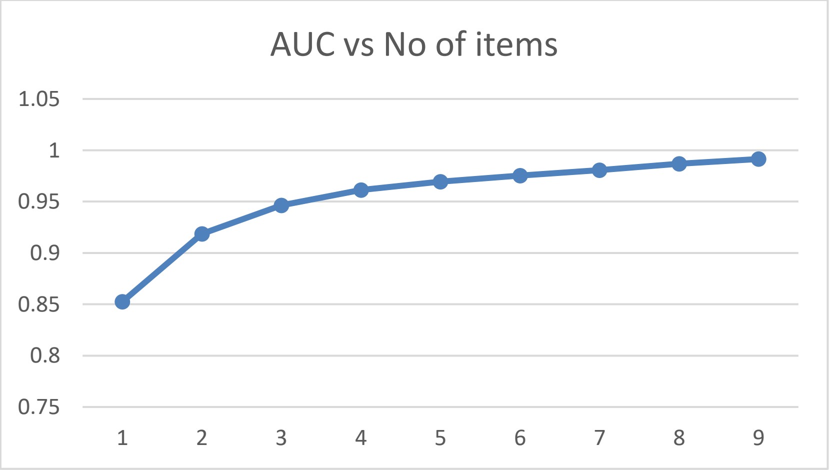
Graph A-5: AUC-ROC Curve for KNN



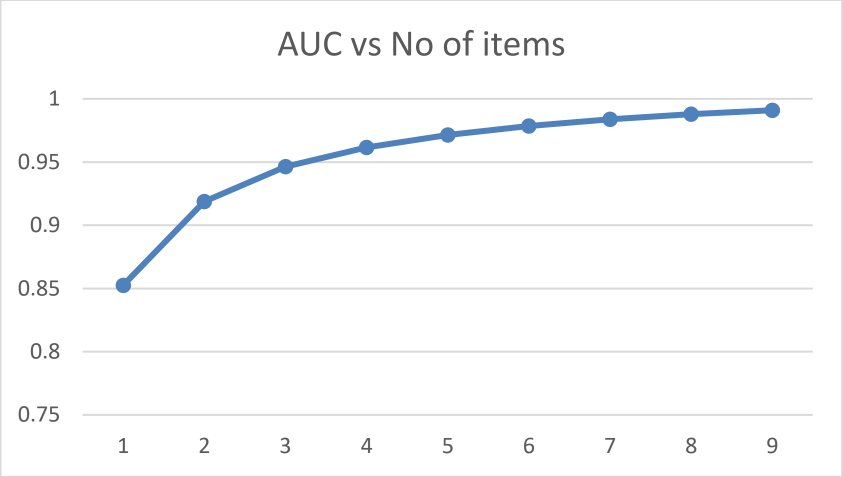
Graph A-6: AUC-ROC Curve for LightGBM



Graph A-7: AUC-ROC Curve for MLP



Graph A-8: AUC-ROC Curve for Random Forest



Graph A-9: AUC-ROC Curve for XGBoost

Appendix B: Tables

Table B-1: AUC Scores for Top Performers Models

Model	Num Questions	Selected Features	AUC
XGBoost	3	(Q1, Q22, Q39)	0.957014672
LightGBM	3	(Q1, Q22, Q39)	0.957009879
MLP	3	(Q1, Q22, Q39)	0.956818177

Table B-2: Best AUC and F1 Scores for Optimized Models

Model	Best AUC	Best F1 Score
XGBoost	1	0.9925
MLP	1	1
Gradient Boosting	1	0.9934
LightGBM	0.9999	0.9918
Random Forest	0.9995	0.9865
AdaBoost	1	0.9957
Decision Tree	0.9740	0.9404
KNN	0.9984	0.9769
SVM	1	1

Table B-3: Comparison of validation accuracies of the best models trained on combinations of 10 questions, averaged over 9 combinations, using default hyperparameters, without demographics

Model	Mean AUC	AUC CI Lower	AUC CI Upper	Range of AUC CI	Mean F1	F1 CI Lower	F1 CI Upper	Range of F1 CI
LightGBM	**96.69%**	96.57%	96.81%	0.24%	<1.0e-04	90.38%	90.21%	90.54%
XGBoost	96.68%	96.55%	96.80%	0.24%	<1.0e-04	90.44%	90.29%	90.60%
MLP	96.65%	96.54%	96.77%	0.23%	<1.0e-04	90.24%	90.10%	90.40%
Gradient Boosting	96.65%	96.53%	96.77%	0.24%	<1.0e-04	90.29%	90.13%	90.44%
AdaBoost	96.58%	96.46%	96.70%	0.24%	<1.0e-04	90.06%	89.89%	90.21%
Random Forest	96.52%	96.41%	96.63%	0.22%	<1.0e-04	**90.49%**	90.32%	90.64%
Decision Tree	95.83%	95.73%	95.92%	**0.20%**	<1.0e-04	90.25%	90.10%	90.38%
KNN	94.75%	94.60%	94.90%	0.30%	<1.0e-04	89.38%	89.15%	89.59%
SVM	94.71%	94.51%	94.89%	0.38%	<1.0e-04	90.26%	90.08%	90.42%