2026-04

The Hidden Complexity of Mental Health: Multi-Entropy Analysis of Response Patterns in Depression Severity Assessment

Siddika, Ayesha

IUB

The Hidden Complexity of Mental Health: Multi-Entropy Analysis of Response Patterns in Depression Severity Assessment

Prepared by

Ayesha Siddika
ID: 2221891
2221891@iub.edu.bd

Maria Jahan Noon
ID: 2222637
2222637@iub.edu.bd

Ahnaf Atif Rafi
ID: 2130352
2130352@iub.edu.bd

Supervised by

Sanzar Adnan Alam

Lecturer A, Department of CSE

Department of Computer Science and Engineering

Independent University, Bangladesh

April 2026

Attestation

We are aware of the fact that plagiarism is strictly prohibited and that it conflicts with the rules and regulations of Independent University, Bangladesh. We hereby guarantee the authenticity of this work. Any source, concept, figure, or copyrighted material used from other authors has been properly cited following accepted academic standards.

Ayesha Siddika	Signature: ....
Ahnaf Atif Rafi	Signature: ....
Maria Jahan Noon	Signature: ....

Date: ....

Acknowledgement

We would like to express our sincere gratitude to our respected supervisor, Sanzar Adnan Alam, Lecturer A, Department of Computer Science and Engineering, Independent University, Bangladesh, for his guidance, encouragement, and continuous support throughout this research. His valuable suggestions and thoughtful feedback greatly improved the quality of this work.

We are also grateful to CCDS (Center for Computational & Data Sciences), Independent University, Bangladesh, for providing an inspiring academic environment and research support. Finally, we express our heartfelt appreciation to our teachers, peers, and families for their encouragement during the completion of this report.

Evaluation Committee

Supervision Panel

Supervisor	Co-supervisor
....	....

Examiners

Internal Examiner 1	Internal Examiner 2
....	....

External Examiner

Office Use

Center Director	Head of the Department
....	....
Director, Graduate Studies, Research and Industry Relations	Library
....	....

Abstract

The Patient Health Questionnaire-9 (PHQ-9) is generally used to determine the level of depression, and clinical decisions are usually made based on the overall score. Nevertheless, there can be significant variation in the response pattern of individuals who have the same score, indicating that there are variations in the organization of symptoms. To overcome this drawback, this paper presents an entropy-based computation system that describes PHQ-9 responses in a way that is not limited by a score measure.

Responses to individual items are fused with information-theoretic quantities, such as Shannon entropy, sample entropy, permutation entropy, and multiscale entropy, to form a more detailed feature space to classify depression severity. The proposed framework significantly increased the accuracy of the sum-score baseline in fivefold stratified cross-validation, with Random Forest improving from 81.2 percent to 99.2 percent. Statistical analysis also shows that the relationship between response entropy and the severity of depression is not linear, where entropy is highest around moderate depression and reduced in severe depression. ROC-based evaluation further shows that entropy has meaningful screening utility for severe depression detection, achieving an AUC of 0.852. Using Youden's J statistic, the optimal entropy threshold was identified as 0.439, with a sensitivity of 0.771 and specificity of 0.793. These results suggest that response-pattern complexity can act as a complementary digital marker for depression severity assessment.

Keywords: Depression classification, Digital biomarkers, Mental health assessment, PHQ-9, Response pattern analysis.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Depression is one of the most common mental health disorders and is a major contributor to disability worldwide. Traditional screening tools such as PHQ-9 remain clinically valuable because they offer a simple and validated way to assess the severity of symptoms. However, individuals with the same total PHQ-9 score may have very different response patterns at the item-level. This raises the possibility that the patterns of symptoms and complexity have important information that is not represented by the sum score.

In this report, we examine how entropy-based modeling can capture variability, inconsistency, and organization in responses to PHQ-9 items. By combining item-level responses, entropy measures, statistical validation, and machine learning, the study explores where a finer distinction severity can be achieved.

1.1 Background of the Study

In recent times, mental health has been an important focus area because it affects a person's emotional well-being. Being healthy does not simply imply that you are not ill. It is also a sign that you are able to cope with stress as well as relate with others and do what you have to. As all the changes of technology and the changes of society are taking place, a lot of changes are affecting the mental health of people including the youth.

There is a problem with mental health issues. Researchers believe that methods based on facts and numbers are extremely valuable in comprehending mental health problems. This holds particularly so in the case of phenotyping and computational psychiatry

One of the prevalent mental health issues in the world is depression. It impacts people of every level. Feeling sad, tired are some of the things that people who are depressed tend to feel. Difficulties in thinking. The other disease is depression. Individuals experiencing depression may react or display various symptoms. This leaves it difficult to determine the ones who are depressed and those who are not. The common methods of testing, on depression, are not so good. They are overly diminutive.

With an increase in stress levels, time in school and in front of screens, people need to improve methods of mental health testings. The measures of mental health are highly valued. We must be careful about our health

1.1.1 Mental Health and Depression

It has to be capable of thinking and feeling well to be able to solve the problems. You can hardly do them when you are depressed. Depression renders us with difficulty in getting things done and in associating with other people. Although depression has been extensively researched by people, it remains difficult to comprehend. Depression is difficult to calculate and categorize under.

The way in which depression is typically examined in physicians offices and research centers is through tests that require the individual to answer how severe their symptoms are [1, 2]. These tests consider depression as identical, to everyone, which is not the case. Dissimilar individuals are affected by depression differently. This means that there is a need to consider a more sophisticated method that can account for the variability of symptoms and other information about mental health disorders [14].

1.1.2 Modern Lifestyle and Mental Health Challenges

Increasing life pressures have played an important role in the increase in mental health problems. Academic and workplace pressures, isolation from social challenges, and a over-reliance on social networks have created a climate in which stress is increasing [8]. These challenges are particularly in students and young people who are more vulnerable to lifestyle changes and digital influences [9].

However, technological advancement has provided novel possibilities for tracking mental health using digital biomakers and behavior data. these approaches enable a real-time and quantitative understanding of mental states, rather than the qualitative approach clinical observations. Consequently, there has been an increasing interest in the application of techniques in mental health science to improve the treatment of psychological disorders. [27, 28].

1.1.3 Need for Screening Tools and Limitations of PHQ-9

In the face of the increasing rate of depression, screening instruments are needed to identify and treat depression. The Patient Healthcare Questionnaire (PHQ-9) is one of the most frequently used measures of the severity of depression due to its brevity and validated structure [1]. It assists individuals to scale the frequency of the symptoms with a scale. This is why it is applicable in clinic and research studies.

The PHQ-9 surveys the score to determine the severity of depression.. This may not give the complete picture. An example of two individuals with the score may have varying symptoms. This may indicate that they are mentally ill. Scientists are currently seeking methods of comprehending the state of mental health. They desire to see more, than scores. One of these methods is through the use of something known as entropy analysis. This contributes to the observation of the trends of the symptoms. [2, 3, 4, 7].

1.2 Problem Statement

Today depression is an issue to many. It influences the way individuals communicate, their moods and motivation, concentration and general living. Simple scoring methods are still many because with the help of tests one can find out whether he/she is depressed or not. The Patient Health Questionnaire PHQ-9 is commonly employed in assessing the severity of badness of someones depression, summing up his or her responses to nine items. This is an easy way of doing things. It assumes that all the scores have the same meaning and this may not be true [1].

Depression has such an effect on individuals. Symptoms can be indicated in a PHQ-9 score, such as emotional, thinking and behavioral problems. This is in the sense that a clear picture may not be seen even by simply looking at the score. Symptoms can be apparent in the PHQ-9 score, which can be misinterpreted. The frequent scoring system may overlook the specifics of response in the individuals.

Mental health records are sometimes unruly and may not be comprehended. Calculating the total score may not be precise. We must have methods to examine data that consider the description of not only the overall score. This can assist us to get a clearer picture of depression [12, 14].

Key Issues Identified

1. Depression screening may overemphasize severity scores at the expense of response patterns.
2. PHQ-9 scores may mask different patterns of symptoms.
3. Conventional approaches yield little information on variability and irregularity.
4. Nonlinear patterns may be present in mental illnesses like depression.
5. Conventional screening may not make full use of item-level information.
6. There is increasing need for more sensitive, data-driven methods.
7. New methods, such as entropy methods, may uncover complexity within sum scores [2, 3, 4, 7].

1.3 Research Objectives

The purpose of this research is to investigate a more holistic method of accessing the severity of depression by extending the conventional scoring approach based on the sum of PHQ-9 responses. Rather than using sum scores, this study seeks to model the response patterns to individual items and quantify their complexity through entropy-based measures. By combining several entropy measures with statistical techniques, the research aims to develop an improved feature set that accounts for intensity and variability. In addition, machine learning algorithms are applied to assess the benefits of this representation for classification compared to conventional methods. In addition, the study explores the nonlinearity between the complexity of responses and the severity of the illness, the discovery of potential subgroups using pattern discovery methods, and explores the potential of having an entropy threshold to better detect severe depression.

1.4 Significance of the Study

This research is as important as it overcomes a key limitation in traditional depression assessment. the severity of depression is typically measured in clinical research and assessed with the total score of PHQ-9, which provides a quick and reliable screening approach [1]. However, using the score alone can miss important differences in discrepancies in the distribution of symptoms in the items. Depression is a complex disorder, individuals with the same sum score can have different response patterns and the structure of the underlying symptoms [12, 14]. This study, which looks at individual items and variation in structure, helps to paint a more accurate picture of depression assessment.

1.4.1 Methodological Significance

The study is important because it uses a way to look at mental health. This new way is called an entropy-based analysis. It helps us to understand how people behave in situations. We can use something called Shannon entropy sample entropy, permutation entropy, and multiscale entropy to measure how people behave. These things help us see how varied and complicated the behavior of people can be [2, 3, 4, 7].

This proves quite useful in instances whereby we would wish to learn about health. Humans do not operate in manners that would be easy to comprehend and by looking at uncomplicated figures we will not be doing anything to comprehend them. These measures and machine learning can help us learn more about the issue of mental health than was possible before using this study of mental health[12]. It is a newer approach to thinking about mental well-being which employs computers and data to assist us make decisions [5, 8, 9].

1.4.2 Practical and Academic Significance

This study is important in two aspects. One is through our activities in clinics and the other is through the larger scientific discourse. The research assists us to filter off the severity of depression that is more than we are capable of doing today. Individuals with PHQ-9 differentials can differ in vital aspects. This paper demonstrates that it is ineffective to look at the score. There must be real differences, between them, we must look closer. Looking outside of the overall score can help us locate useful information. Such a depth of information might prove mighty helpful to those professionals that are attempting to determine the severity of depression in more than their scores would otherwise reflect.

A long-standing basic issue in computational psychiatry and digital mental health is how to apply data-driven tools in clinical care without injuring the proven decades-old tools. This study does not even strive to resolve that tension by taking a specific position. Rather, it provides that the response of one to each question provides information not implying that they must cease PHQ-9 usage. It warrants more attention at the expense of concentration, more concentration [5, 8, 9].

1.5 Scope of the Study

This paper argues that there is no single score for severity of depression. The sample examines the responses of university students to the PHQ-9 in an age group with high occurrence but low detection of depression. The analysis does not only focus on the score but rather on their individual answer patterns to explore what may be learned from them.

However, the key observation is that there is a pattern about how human answer these questions and this patterm contains information. Information is obtained by measuring the regularity of the time series samples through Shannon entropy Sample entropy, Permutation entropy and Multiscale entropy. These metrics allow to develop an understanding of how complex and diverse the responses are.

Another person could get that score also but they might answer those questions very differently and you would be left wondering about the differences in symptoms, since there are no one to one relation. The research combines these measures with machine learning and attempts to differentiate. The problem is to see whether the severity of depression can also be classified accurately, not replace the PHQ-9. By extracting additional knowledge from it. The PHQ-9 is a tool. However, the study wishes to explore whether it can be used more granularly.

This study is based on a set of data, methods and contextual factors When reading the results, just be sure to remember that. This data will form the basis of the results. The sample consists in university students. Your validation method is a stratified cross-validation on the data set which means this outcome is justifiable but must be validated over external populations or additional other groups.

This work should not be used instead of clinical judgement and does not replace PHQ9 screening. It is designed to be an overlay as opposed to a supplement for an established screening practice. It is also longitudinal not cross-sectional study. It is confined to the data available through clinical history, without any access to external data such as psychological indicators and behavioral patterns. These are limitations, not mistakes. This research represents a advancement on the path, however not yet here is your definitive bearer of low state of mind in clinic. Its worth is that it expands avenues to future study rather than answers in itself. [1, 2, 3, 4, 7, 10, 11, 12, 13, 14].

1.6 Limitations of the Study

There are some limitations of this work that need addressing. Consider who the data actually represents first. In this case the subjects of this study are university students. This group is very nice to deal with. It lacks diversity in demographics and clinical experience. The findings from this group may not be relevant to different groups or those receiving treatment from psychiatrists. The patterns we detected are quite unique to this sample of university students.

The second limit is related to the data. The PHQ-9 questions(17) are self-reported, which means they are subjective. The answers may depend on how they feel answering the questions or whether they are being honest. That is not to say self-reporting is a technique, but that the data has errors embedded within it that are difficult to parse out from fact.[1, 14].

Another limitation is how was tested the framework. The numbers that we present is from prediction, which is a good approach but it needs to be 'tested' against unseen data. The positive outcomes we received are an indication, but we must be cautious in interpreting them. To know for certain that it works elsewhere, we must test the framework with data.

This research has limitations as a result of the type of data it examines. The responses to a survey, taken at a specific moment in time can provide only part of the whole picture. Depression is not a staticr mentalstate it runs on a timeline and to take hold of the depression we need both time-series data as well as physiology signals or behavioral-based data. Lack of such information did not represent a failure of the study. That is a maximum of what you can do now.[8, 9].

Taking all those restrictions into account, the results are great. Keep to sound methods but not be presumptuous they apply to all. The next step is to apply the framework in practice and with different groups.

Key Limitations

- The data we are looking at is about students.
- We got our information from a questionnaire that the students filled out themselves so we are relying on what they said about their symptoms. This way of doing things

is simple and lots of people use it when they study health but it is not perfect because it is based on what the students think not on what a doctor might say.

- We only looked at the data we had we did not check it against any data to see if it was correct.
- Our findings are based on data from one point in time so we do not know how the symptoms of depression changed over time for each student.
- This study only looks at what the students said on the questionnaire it does not look at things like what their bodies are doing or how they are behaving.
- What we found out is useful because it helps us understand more about depression but it is not enough, on its own to diagnose someone it should be used with tools that doctors already use to help people with depression.

1.7 Organization of the Report

The report is divided into seven parts. Each part adds to what was said. This makes the report easy to understand. The studys main point is explained in an simple way. The report is. Easy to follow. The seven parts of the report work together to make the studys argument clear. Chapter 1 introduces the background of the study, the problem statement, the research objective, the significance of the study, the scope, and the limitations. Chapter 2 presents a review of the literature. Chapter 3 describes the methodology of the study, including data set preparation, preprocessing, entropy-based feature extraction, machine learning models, statistical validation, clustering methods, and threshold optimization. Chapter 4 provides the results and analysis of the proposed framework. Chapter 5 presents selected case studies from real-world digital mental health research. Chapter 6 discusses the broader meaning and implications of the findings. Finally, Chapter 7 concludes the report by summarizing the major findings, acknowledging the limitations, and suggesting possible directions for future work.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the current literature related to the assessment of depression, PHQ-9-based screening, entropy-based and machine-learning-based behavioral analysis, and machine-learning-based mental health research. This review defines the theoretical and empirical basis of the current study by reviewing the measurement of depression as conventionally done, the limitations of previous score-based systems, and how new computational approaches have begun to overcome these problems. As the present study aims at the complexity of response patterns in PHQ-9 data, it is necessary to understand the clinical applicability of standardized depression screening instruments, as well as the increasing significance of data-driven methods in digital mental health [1, 5, 9].

The literature also indicates that despite the wide acceptance of instruments such as PHQ-9 in depression screening, their interpretation has largely focused on the total score and therefore might not identify key variations in the way an individual displays symptoms [14]. Concurrently, advances in areas such as information theory, computational psychiatry, and machine learning have dramatically increased the scope of what is achievable when utilizing psychological data. The entropy-based approaches offer something that the conventional methods do not add, they provide a means of modelling variability, irregularity and the underlying complexity in how people respond to rather than what they report. For its part, machine learning has been especially able to reveal patterns structures in data potentially obscured or flattened by traditional statistical methods [2, 3, 4, 7, 12, 15, 16, 19, 20, 21]. The current chapter goes beyond a review of what has previously been covered against this background. Not only that, but it carefully outlines the borders of existing work such a way that it highlights where the holes exist - and it is those very holes which create both the context in which this new model for assessing depression severity (with added entropy) can even have existence but also context in which its motivational significance does actually live.

2.2 Depression and Mental Health Assessment

In both modern healthcare and psychological research, assessment of mental health has become integrated. Mental health is not that simple. They dont always fall into just one category? You cannot test only one, to diagnose what is wrong. Mental diseases are by no means universal, every one has diverse problems. They are dependent on the fate of each individual. Doctors and researchers therefore combine the two to research health. They analyze what people talk about with their symptoms. Investigation of mental health is some thing that enormously regarded. Mental health is complex and variable. They are profoundly molded by what individual people go through. Psychological health assessments assist doctors and researchers in understanding these issues. They employ health consultation to better understand, about mental illness matters. Also behavioral observation, standardized rating scales, and professional interpretation to build an understanding of what a person is going through — each source adding something the others alone cannot fully provide. Selection and design of assessment tools are very important because quality screening affects early detection, treatment decisions, and ongoing support. This includes assessment tools. Quality matters in screening: Screening quality is a vital component for problem discovery and good treatment plans. Thats why the design and choice of assessment tools is so important, for treatment planning and long-term follow-up. Evaluation is particularly important in depression because the disease can affect emotional state, cognitive ability, energy level, sleep, appetite, and social functioning simultaneously [1, 5].

Mental health assessment has developed over time from less structured clinical assessment to more standardized and evidence-based processes. Structured screening instruments help researchers and clinicians quantify depressive symptoms in a more consistent manner. However, despite the reliability gained through standardization, this has also encouraged the use of simplified scoring methods in which sophisticated symptoms experiences are often condensed into one total score. Consequently, contemporary evaluation techniques are appropriate for the screening, yet they might still fail to represent significant variations in symptom expression across individuals [12, 14].

2.2.1 Depression as a Heterogeneous Mental Health Condition

Depression is a prevalent and devastating mental health disorder across the world. It is generally marked by persistent sadness, loss of interest in daily activities, loss of motivation, despair, cognitive difficulties, and disruption of physical and emotional functioning. Depression does not, however, present itself in one uniform way. Some individuals may show more emotional symptoms such as sadness or guilt, while others may show more cognitive symptoms such as poor concentration or somatic symptoms such as fatigue and sleep disturbance. This heterogeneity makes depression a complex condition [1, 12].

The heterogeneity of depression poses a significant assessment problem. Two individuals can be assigned to the same severity designation, but the distribution of their symptoms can vary significantly. Two individuals could have scored the same number, the sum of the package, in the PHQ-9 test, but did so in radically different ways. Someone could feel kinda bad about a lot of things but not really all that bad about any one thing. The other, far/far less likely to feel bad about things, but definitely not leaving the rest of the world with people feeling bad about them. And the point to which they arrive the same. However, the truth of what each individual is in fact going through might be very different. Indeed, it is not an odd kind of conduct that takes place with figures. It is noteworthy because it tells us about the specific way in which the depression appears to be in real life. The more specifics, then, that we could look into, in order to determine how bad the symptoms of someoneself are or were in particular (rather than how bad they feel in general). Rather we need to look at which symptoms are problematic and how (and to what extent) such symptoms can impair the person.[14].

2.2.2 Importance of Standardized Assessment Tools

Depression shows up differently in each person. That's why standardized tools for assessing depression have become very important in both practice and psychiatric research. These tools are valuable because they help evaluate symptoms in a way. They use a structure that can be repeated which allows findings to be compared across different people, institutions and research settings.[1]

These tools protect against the variability that can come with observation. Experienced clinicians can have different opinions but structured questionnaires help reduce this subjectivity by using clear criteria. This combination of reliability, consistency and usefulness across settings has made standardized tools essential in clinical screening and large-scale research.

Standardized tools have another benefit: they produce data that can be easily analyzed. When symptoms are recorded in a format the data becomes quantifiable, sortable and open to analysis. This is much harder to do with clinical notes. Researchers can examine how reliable the data is, categorize the severity of symptoms identify patterns across populations and apply methods more efficiently.

This is especially important today. With the growing use of phenotyping, machine learning and behavioral analytics in mental health research the need, for clean and structured data has increased. Standardized assessment tools are the foundation of all this work. They connect the methods of measuring psychological conditions with the new more advanced computational methods that are changing how we understand mental health conditions.[5, 8, 9].

Table 2.1: Common Approaches to Depression Assessment

Assessment Type	**Example**	**Strength**	**Limitation**
Self-report scale	PHQ-9	Simple, fast, scalable	May overlook symptom structure
Clinician-rated scale	HAM-D / MADRS	Rich clinical interpretation	Time-consuming
Diagnostic interview	DSM-based interview	Deep clinical insight	Less practical for large populations

2.2.3 Common Methods Used in Depression Assessment

When we talk about measuring depression there are three ways to do it. Each of these ways has a purpose. The common way is to use self-report questionnaires. These questionnaires are easy to give to people they do not cost a lot of money. They can be used with a large number of people. They ask people to describe how they are feeling. They have set answers to choose from. This makes it easy to look at the answers and compare them.

Clinician-rated scales are different. They do not just rely on what the person says about themselves. Instead they use a trained professional to help figure out how the person is doing. This gives a complete picture of the persons condition. However it takes time and resources so it is not always possible to use this method.

Diagnostic interviews are the detailed way to measure depression. They give a lot of information that the other methods cannot. That is why they are used to make diagnoses. However they take a lot of time. Require special training and planning. This makes them hard to use when you need to screen a lot of people[1, 14]

Among these methods self-report scales are very important in health and research. They are useful because they can be used with a lot of people quickly. They also give answers that're consistent and can be compared. However there is a problem with the way these scales are usually used.

The problem is that all the answers are added up to give one score. This is easy to do. It does not tell us how the person got that score. Two people can have the score but they can have very different answers. One person might have a bit of distress in many areas while the other person has a lot of distress in just a few areas. The total score does not show this difference.. It is this difference that can give us important information, about the persons depression.[12, 14]

2.2.4 Emerging Shift Toward Data-Driven Mental Health Evaluation

Mental health research is changing the way it looks at things. Lately it has been moving away from using traditional scoring methods. Instead its using approaches that're more based on data and computers. This change is happening because of developments in digital mental health, computational psychiatry and digital biomarkers.

These areas are all interested in understanding health conditions by looking at how people behave. They want to use this information to help than just relying on summary statistics. The idea behind these directions is that the data collected from standard assessment tools has more to it than traditional scoring methods can show. This change is especially important for depression research.[5, 8, 9]

Depression is a condition that doesn't behave in simple ways. Its symptoms can be different from person to person. Can interact with each other in complicated ways. Traditional scoring methods can't always capture this. New analytical methods can help us look deeper into the data.

They can show us how individual symptoms contribute to the score. They can also help us see if there are subgroups of people with depression that have patterns. This study is part of a movement in research. It brings together approaches and traditional psychometric tools. The idea is that these approaches are more useful when used together. Each one can provide something that the other can't. Mental health research and depression research are areas where this approach can make a difference. Depression research benefits, from this combined approach. Mental health research also benefits from this perspective.[12]

2.3 PHQ-9 is a Depression Screening Tool

One of the self-report tools that have been used widely to identify and measure the level of depression is the PHQ-9. The number of instruments in mental health research that have reached the same adoption rate as PHQ-9 has was very small. It has been developed as a concise yet clinically rigorous tool that is based on known diagnostic cutoffs of depressed disorders and has found a home in both healthcare environments and in the research world in general, valued for the balance that it creates between simplicity and meaningful measurement -which is valued at all costs even in today's world of sophisticated and cutting-edge technology and globalization. The reason it is especially effective when used in large-scale screening is because it does not require a lot of time and effort on the part of the respondent, and yet, it does generate structured and quantifiable information that may be used to classify the severity and draw comparisons between an individual and a group.[1]

The best feature of the PHQ-9 that makes sense in a computational process is the manner in which it addresses subjective experience. Instead of leaving the description of the symptoms open-ended, it takes the respondents through a series of predetermined depressive symptoms and asks them to indicate the frequency to which each of the symptoms has been present during one of the recent periods. This arrangement does something important, it transfigures what is by nature personal and variable, into a form, which is systematic, consistent and readable across different persons and circumstances. Since the responses are recorded in numerical terms, the instrument also provides a natural access to statistic analysis, psychometric assessment, and computational modeling that will enable researchers with a solid ground on which more complex approaches can be constructed based on the results of the preceding steps in the study's methodology [1, 16, 17].

2.3.1 Structure and Item Composition of PHQ-9

There are nine items in the PHQ-9. Each item outlines one symptom of depression. These are already symptoms that physicians and scientists have been familiar with for some time. But there is a lot that the PHQ-9 covers— Its into how people feel, like sad all the time or dont care about anything anymore. It is also concerned with issues of thinking such as when people have difficulty concentrating. It includes questions that examine behavior change and physical symptoms (for instance, when people cannot sleep, are lethargic or eat more or less than their usual). We even examine issues like self-hatred and suicidal ideation [1]. It does this intentionally via the PHQ-9. You wondered how this could be, because depression is different, for everyone. Instead of asking just one general question, the PHQ-9 can get a more accurate image of what is going on in each individual by asking for symptoms.

Table 2.2: PHQ-9 Items and Symptom Domains

PHQ-9 Item	**Symptom Focus**
PHQ-1	Little interest or pleasure in doing things
PHQ-2	Feeling down, depressed, or hopeless
PHQ-3	Trouble falling or staying asleep, or sleeping too much
PHQ-4	Feeling tired or having little energy
PHQ-5	Poor appetite or overeating
PHQ-6	Feeling bad about yourself or feeling like a failure
PHQ-7	Trouble concentrating on things
PHQ-8	Moving or speaking slowly, or being restless
PHQ-9	Thoughts of self-harm or being better off dead

2.3.2 Likert Scale Response Format

All PHQ-9 items are rated on a four-point Likert scale indicating how often a symptom is experienced during a specific recent period. The answer types will be Ususually about 0 not at all, 1 several days, 2 more than half the days, and 3 nearly every day. The form is practical since it enables subjective mental experiences to be encoded in a convenient and quantifiable manner. As everything is represented by a common numeric approach, the data can be summarized, compared and modeled more effectively.[1]

2.3.3 Scoring and Severity Interpretation

The individual response values for each of the nine items are summed to produce a total score ranging from 0–27 once all nine items have been completed [1]. This mask is then transformed to the corresponding set of predefined severity intervals and provides for clinicians and researchers an immediate, quantitative and easy reference point on how well a subject currently classifies according to depression severity at the moment of assessment.

Table 2.3: Standard PHQ-9 Severity Interpretation

Total Score Range	Severity Level
5–9	Mild
10–14	Moderate
15–19	Moderately Severe
20–27	Severe

2.3.4 Strengths and Limitations of the PHQ-9

The PHQ-9 is arguably the most established clinical tool with decades of psychometric validation in various populations and briefer length, ease of use and wide availability make it more than well suited for routine screening or large-scale research [16, 17, 18, 19, 20]. Moreover, its item-level responses are structured and therefore can be repurposed for many applications beyond mere classification.

But there is a tension at the center of how the tool is typically used. While necessitating efficiency in the form of one aggregate score out of nine different symptom responses, this approach necessarily flattens differences that may be clinically meaningful. Two persons may have arrived at the same totals whilst endorsing completely different symptoms or the same symptoms with very different severity [14]. But under a sum-score framework, those differences vanish. This concern becomes more urgent as our understanding of depression evolves away from the view that it is a single condition with a discrete and uniform signature towards one recognizing a syndrome that is heterogeneous in its presentation, nonlinear regarding symptom interactions, and dynamic as symptoms change over time. [12].

2.4 Limitations of Sum-Score-Based Assessment

Using a sum-score approach to measure screening for mental states has understandably remained. It is quick, easy to communicate and provides clinicians with an operational shorthand for assessing where a patient is — all of which are relevant in contexts where time and resources are often stretched. For PHQ-9 and other tools, the logic is straightforward: total item responses and read the sum against a pre-defined severity range. This is good enough for a very wide classification.

Where the challenge lies is in what the approach quietly assumes. Ascribing identical clinical meaning to patients with identical total scores means treating the number as if it is a complete and sufficient representation of the underlying condition — an assumption that does not fare well when hinged on such a heterogeneous disorder as depression [12, 14]. The sum-score framework offers no differentiation between two people arriving at identical totals through completely dissimilar symptom combinations.

This points to a larger issue. Depression does not spatially compress into a linear dimensionality. Such that it crosses emotional, cognitive, behavioral and somatic domains simultaneously, and the way those domains interact with each other varies widely among individuals[1] This has the effect of discarding all information about the configuration that gave rise to this number when responses across all nine items are folded into one: the pattern of endorsement across items (like- whether symptoms tend to appear together or separately, and in what proportion) is lost entirely. What you are left with is a number that communicates something about severity overall, little as to the structural details of the experience it represents.

Illustrative Example

- Person A: 2, 2, 2, 2, 1, 1, 1, 1, 0
- Person B: 3, 0, 3, 0, 3, 0, 2, 1, 0

Both patterns can have the total score but they do not show the same organization of symptoms. Person A has their burden spread out evenly. On the hand Person B has severe problems, in a few specific areas. A system that only looks at the score might think they are the same. However if we look at the patterns we can see that the way they responded is actually different. This is because the internal structure of their responses is not the [14].

Table 2.4: Why Sum-Score Assessment Can Be Insufficient

Limitation	Explanation
Loss of item-level detail	Individual symptom differences are hidden within one total score
Assumption of equivalence	The same score is treated as the same severity profile
Reduced interpretability	Symptom organization and variability are not reflected
Inability to capture complexity	Nonlinear and irregular response behavior is ignored
Risk of overlooking subtypes	Distinct response patterns may be grouped together

2.5 Nonlinear and Information-Theoretic Perspectives in Mental Health Research

An expanding literature in computational psychiatry and digital mental health has started to challenge the paradigm of linear and aggregate methods being sufficient for modeling psychological phenomena [5, 9, 12]. Human response — especially when it comes to mental health — tends to be patchy, context-dependent and most likely influenced by factors that summary statistics are ill-designed for picking up. As a result recognition to study behavioral and psychological data using information-theoretic and nonlinear frameworks are now considered as alternative lens.

Entropy has been one more active notion in this area [2]. While traditional thinking focuses on symptom severity, entropy-oriented thinking asks new kinds of questions — how are those symptoms ordered, how reproducible are they, and what is the variability in an individual response profile. These are not trivial distinctions. This type of structural perspective provides something measurement-based measures of magnitude cannot achieve in the context of depression, which captures very heterogeneous symptom expression within and across individuals over time [12, 13].

What this literature ultimately delivers to the current study is a guiding principle: that behavioral and psychological data have structure within their summary scores, and that such latent information may represent the most clinically meaningful source of variance.

2.6 Machine Learning in Depression Assessment

Machine learning is really important for health research and it is easy to see why. The usual ways of looking at statistics need us to have some ideas about what we're looking for before we start. But machine learning models do things differently. They look at the data. Find patterns and connections that we might not have thought of before. This is very useful when we are studying depression because people can have different symptoms and these symptoms can affect each other in complicated ways. So machine learning is very good at understanding these relationships.[15].

We have seen new ways that machine learning can be used in recent years. Machine learning can be used to classify symptoms of depression and predict how severe they will be. It can also be used to monitor what people do and find ways to measure depression using digital tools. Machine learning can even look at data from things like smartphones to find patterns that might tell us something about depression. This shows that machine learning can be used in different ways to learn more, about depression.[8, 9, 27, 28, 29, 30, 31, 32, 33, 34, 35].

What is really important in this work is that computational models do a job when they get good information. The features that are put into these models need to be really good. Well made. They should be better, than using raw scores. The kind of information that is put into the model is what makes the results good or not. Computational models are able to give us more when the featuresre richer and more carefully made.

Table 2.5: Selected Applications of Machine Learning in Depression Research

Study Area	**Example Focus**	**Relevance**
Digital biomarkers	Passive sensor data	Behavioral signals beyond questionnaires
Symptom prediction	Clinical note or response analysis	Severity estimation
Screening systems	Smartphone-based detection	Scalable mental health assessment
Multimodal analysis	Combined data sources	More robust classification

2.7 Related Studies and Research Gap

There are three ideas in the existing research that are really important to this study. The first idea is from research on tests and PHQ-9 studies, which have shown that checking for depression in a structured way is very useful and that just looking at the total score is not enough to understand the results. This is what the studies in references [14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25] have found.

The second idea comes from research on dynamics and entropy-based methods. This research has shown that looking at how complex behavioral responses are can give us an understanding of how depression works and how it is different for each person. We can see this in the work done in references[12, 13].

The third idea is from the research on machine learning in health. This research has shown that computers can be used to find patterns in symptom data that we cannot see otherwise. This can help us diagnose depression better than we can with methods. The studies, in references[8, 9, 15, 27, 28, 29, 30, 31, 32, 33, 34, 35]have all shown this.

There is still a problem where these three things meet. The PHQ-9 is used a lot. When we try to figure out how depressed someone is we mostly look at the total score. People have studied entropy in terms of behavior and mental health. We have not really used it to help us understand PHQ-9 scores in a way that compares the old way of scoring to a new way that uses entropy.

We know that machine learning can be helpful, with health but not many people have tried to use PHQ-9 answers, entropy measures and predictive modeling together in a way that makes sense and is easy to understand [12, 14].

2.8 Chapter Summary

The thing is, even though we have made progress in these three areas there is still a gap where they all come together. The PHQ-9 is still one of the used tools to screen for depression in clinics and research but most of the time people just look at the total score and that is it. Methods that use entropy have shown a lot of promise in studying behavior and mental health. We have not really used them in a systematic way to classify how severe depression is based on the PHQ-9. We have not really compared what the traditional way of scoring tells us to what we can learn from using entropy. Machine learning has also been very useful in health but there are not many studies that bring together the answers to each question on the PHQ-9 different measures of entropy and a way to predict what will happen all in one framework that makes sense.

This is where the present study comes in. The PHQ-9 and entropy and machine learning all meet here. The problem is not about the technology. It is about how we think about these three areas of research. We tend to think of them as separate than, as working together and that is what this study is trying to change. The PHQ-9 is still an important tool and entropy and machine learning are still very useful but we need to bring them together in a better way.

Chapter 3

Methodology

3.1 Introduction

This chapter explains how we did our study. We made some decisions followed steps and had reasons for doing so. Our main goal was to go beyond a simple score and create a better understanding of how people express depressive symptoms in their PHQ-9 responses. To do this we built a step-, by-step computer process. We developed it in stages making sure each step made sense based on the one. The process started with collecting and cleaning up data. Then we used a method based on entropy to extract features. Next we used machine learning to classify things. After that we looked for patterns without being too specific. We also checked our results with statistics. Finally we fine-tuned some settings. This sequence shows what we wanted to achieve with our study and the limitations we had working with one specific data set. The study uses PHQ-9 responses to understand symptoms. We analyzed PHQ-9 responses to identify patterns. The PHQ-9 responses helped us understand how depressive symptoms are expressed.

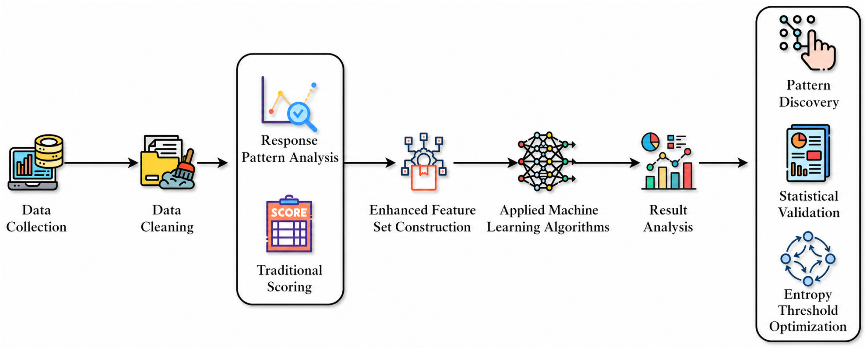

Figure 3.1: The Workflow Diagram

3.2 Research Workflow Overview

The research process has steps. Each step is connected to the step in a logical way. It starts with collecting and preparing the PHQ-9 questionnaire responses. We need to make sure the information is, in a form so we can really understand it. Then we look at the quality and consistency of the PHQ-9 questionnaire responses. We do this before we do anything. It is a check to make sure everything that comes after it is reliable and based on a solid foundation. The PHQ-9 questionnaire responses are very important. We need to get them right.

We have validated data now. We use this data to get different kinds of information about each item. This information is very detailed. It helps us understand the symptoms better than just looking at a total score. We use this information to train machine learning models. The goal of these models is to see how our new method works compared to the way of just adding up scores.

We are also doing some work where we let the computer look for patterns in the data, on its own. We want to see if there are groups of people who answer the questions in ways. These groups might be important for doctors to know about even if we cannot see them when we just add up the scores. At the end of the study we check to make sure our findings are correct and useful. We do this by using statistics and finding the way to use our results in real life. We use symptom complexity and machine learning models to do this. Symptom complexity is a part of our study and machine learning models help us understand it better.

Main Stages of the Workflow

- We start with collecting data and getting the dataset ready.
- PHQ-9 response representation
- We try to find the threshold for the model and see if it is really useful, for doctors and patients.
- Then we do some work to make sure the data is good and we can use it.
- We look at the data to find the parts using things like entropy and statistics.
- Unsupervised pattern discovery
- Statistical validation
- Next we make a model that can learn from the data. We test it to see how well it works.

3.3 Study Design and Dataset

The study is based on numbers and computer work using answers from the PHQ-9 questionnaire. We got 2,039 completed questionnaires from university students at different schools. We collected these questionnaires in a way that kept the students anonymous. We wanted to focus on students because they're a group where depression is a big problem. It is really important to find depression in students. The answers from the questionnaires are the information we used for the entire study. We used this information to look at the details classify the answers, group them together and study the statistics. The PHQ-9 questionnaire answers are the basis, for everything we did in the study.

3.3.1 PHQ-9 Item Structure

But because those nine items also make up the questionnaire we use placed in Inverse order below.:

- PHQ-1: Little interest or pleasure in doing things
- PHQ-2: Feeling down, depressed, or hopeless
- PHQ-3: Trouble falling or staying asleep, or sleeping too much
- PHQ-4: Feeling tired or having little energy
- PHQ-5: Poor appetite or overeating
- PHQ-6: Feeling bad about yourself or like a failure
- PHQ-7: Trouble concentrating on things
- PHQ-8: Moving or speaking slowly, or being restless
- PHQ-9: Thoughts of self-harm or being better off dead

3.3.2 Likert Scale and Response Encoding

The frequency of occurrence of the symptom is rated on a four-point Likert-scale for each item in PHQ-9:

- 0 = Not at all
- 1 = Several days
- 2 = More than half the days
- 3 = Nearly every day

Therefore, in PHQ-9 can be between 0 (no depression symptoms) and 27 (severe depression). This Dataset has Severity Labels of defined based on the general clinical cutoff ranges of PHQ-9 according to standard ranges: Mild (5–9), Moderate(10-14), Moderate (15-19), and Severe (20-27)

3.4 Data Preprocessing and Quality Assurance

Prior to feature extraction or model training, a preprocessing step is applied to the raw PHQ-9 responses that creates clean and uniform data preceding any analysis. People who did not complete at least the first two items are considered as non-completing and excluded from any analyses (response profiles that go to the level described are providing too much uncertainty into further analysis). The gap—this is used when only one or two items are missing. It is filled using the median value recorded for that item across the full dataset — a conservative and commonly accepted approach that preserves the case without distorting the overall distribution.

To see if the questionnaire answers make sense together we check if they are consistent using a measure called Cronbachs alpha. The result is 0.89, which's a good score that shows the nine PHQ-9 items are reliable and measure the same thing for this group of people. This is news because it means we can trust the data and use it for the next steps, which are to extract important features and build models with the PHQ-9 items and the dataset being reliable tools, for our analysis. We feel confident that the PHQ-9 items are working well together to measure what they are supposed to and that gives us a foundation to move forward with our analysis using the dataset.

3.4.1 Response Vector Representation

Each participant's response is modeled as

$$R = (r_1, r_2, \ldots, r_9), \quad r_i \in \{0, 1, 2, 3\}. \tag{3.1}$$

This representation is important because it keeps the order of items in the questionnaire. It helps the study to see how symptoms are arranged. The study can then look at the symptoms one by one than just adding them up, to a total score right away. This way the study can understand the symptoms better.

3.5 Feature Engineering

The main thing this study does is create a set of features called the enhanced feature set. The enhanced feature set combines the answers people give to the PHQ-9 questions with some features that are based on how mixed up the answers are and some statistical features. This study does not just look at the score of the PHQ-9. It looks at each answer, to the PHQ-9 questions. Turns them into a new set of features that can show how bad the symptoms are and how complicated the answers are. The enhanced feature set is used to get an understanding of the PHQ-9.

The final set of features we have is made up of a few things:

- It includes the 9 original PHQ-9 item responses.
- We also have 4 features that are based on entropy and 5 features that are based on dispersion.
- So when we put all of these together the total feature space contains 18 features.

The total feature space is made up of these 18 features.

3.5.1 Entropy-Based Features

The things we look at to measure entropy are the following:

- Shannon Entropy
- Sample Entropy
- Multiscale Entropy
- Permutation Entropy

3.5.2 Additional Statistical Descriptors

The statistical features we look at are the following:

- Variance
- Skewness
- Kurtosis
- Zero-response count
- Extreme-response count

3.6 Entropy-Based Feature Extraction

This study thinks that how bad someones depression is depends on a things. It is not about how bad the symptoms are. It is also about how consistent the symptomsre if they are unpredictable and how they are organized when we look at the answers to the PHQ-9 questions. We are talking about depression severity and the PHQ-9 items here. We want to see how the responses, to the PHQ-9 items are related to depression severity. To quantify these characteristics, four entropy-based measures are computed from each participant's response vector.

3.6.1 Shannon Entropy

Shannon entropy is used to measure how dispersed the responses are across the four Likert categories. It is defined as

$$H_S(R) = -\sum_{k=0}^{3} p(k) \log_2 p(k), \tag{3.2}$$

where $p(k)$ is the probability of observing response value k in the response vector R. To ensure comparability across individuals, the entropy is normalized as

$$H_S^{norm} = \frac{H_S(R)}{\log_2(4)}. \tag{3.3}$$

A lower value indicates more uniform responses, while a higher value indicates greater response diversity.

3.6.2 Sample Entropy

Sample entropy measures the irregularity or unpredictability of the response sequence when treated as a short symbolic time series:

$$SampEn(m, r, N) = -\ln\left(\frac{A_m(r)}{B_m(r)}\right), \tag{3.4}$$

where $m = 2$ and $r = 0.2\sigma$ in the study setting.

3.6.3 Multiscale Entropy

Multiscale entropy extends sample entropy by examining complexity across multiple resolutions. First, coarse-grained sequences are computed:

$$y_j^{(\tau)} = \frac{1}{\tau} \sum_{i=(j-1)\tau+1}^{j\tau} r_i, \tag{3.5}$$

where τ is the scale factor. Sample entropy is computed at each scale, and the average across $\tau = 1, 2, 3$ is

$$MSE = \frac{1}{3} \sum_{\tau=1}^{3} SampEn(y^{(\tau)}). \tag{3.6}$$

3.6.4 Permutation Entropy

Permutation entropy measures the unpredictability of the ordinal arrangement of the response sequence:

$$PeEn(n,d) = -\sum p(\pi)\log_2 p(\pi), \tag{3.7}$$

where $n = 3$ and $d = 1$. The normalized form is

$$PeEn_{norm} = \frac{PeEn(n,d)}{\log_2(n!)}. \tag{3.8}$$

3.6.5 Additional Statistical Descriptors

To complement entropy, five additional statistical descriptors are computed:

$$\sigma^2 = \frac{1}{9}\sum_{i=1}^{9}(r_i - \bar{r})^2, \tag{3.9}$$

$$\gamma_1 = \frac{\frac{1}{9}\sum_{i=1}^{9}(r_i - \bar{r})^3}{\sigma^3}, \tag{3.10}$$

$$\gamma_2 = \frac{\frac{1}{9}\sum_{i=1}^{9}(r_i - \bar{r})^4}{\sigma^4} - 3, \tag{3.11}$$

$$N_0 = \sum_{i=1}^{9} I(r_i = 0), \tag{3.12}$$

$$N_3 = \sum_{i=1}^{9} I(r_i = 3). \tag{3.13}$$

Table 3.1: Summary of Features in the Enhanced Feature Set

Feature Group	**Features**
Raw PHQ-9 features	PHQ-1 to PHQ-9
Entropy features	Shannon, Sample, Multiscale, Permutation
Statistical features	Variance, Skewness, Kurtosis, Zero-count, Extreme-count

3.7 Machine Learning Framework

To evaluate the predictive usefulness of the enhanced feature set, the study compares multiple machine learning models under different feature configurations. Three main configurations are tested:

- PHQ-9 total sum score only
- Item-level PHQ-9 responses
- Enhanced Feature Set (EFS) = item responses + entropy features + statistical descriptors

The classifiers used in the study are Random Forest, Gradient Boosting, XGBoost, Support Vector Machine (RBF kernel), Logistic Regression, and Multilayer Perceptron (ANN).

3.7.1 Evaluation Strategy

All models are evaluated using 5-fold stratified cross-validation, where each fold preserves class representation. The following performance metrics are used: Accuracy, Precision, F1-score, and Cohen's Kappa. The conventional PHQ-9 total score serves as the baseline representation.

3.8 Unsupervised Pattern Discovery

The study goes further than classification. It also looks at the response data to see if there are any groups or types that we cannot see when we just look at how severe something is. To do this we use methods that can find patterns on their own. We want to see if there are any groups, in the response data.

3.8.1 t-SNE Visualization

We use t-SNE to make it easier to see the response patterns. This method helps us take a lot of data and show it in a simpler way. It keeps the details and helps us see if there are any clear groups that stand out. We use t-SNE to see the response data in a way that's easy to understand.

3.8.2 K-Means Clustering

K-Means clustering with $k = 4$ is applied to partition the data into four groups by minimizing within-cluster variance:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2, \tag{3.14}$$

where S_i are clusters and μ_i are their centroids.

3.9 Threshold Optimization and Clinical Utility

To make entropy really work for us as a way to screen things the study finds the entropy threshold using something called Youdens J statistic:

$$J(\tau) = Sensitivity(\tau) + Specificity(\tau) - 1, \tag{3.15}$$

and the optimal threshold is obtained as

$$\tau^* = \arg\max_{\tau} J(\tau). \tag{3.16}$$

3.10 Statistical Validation

The study tries a lot of ways to check the results and see how the model behaves in different situations and with different settings. It does this to make sure the findings are correct and to compare how the model does with levels of severity and, with different features.

3.10.1 One-Way ANOVA

ANOVA is used to test whether entropy values differ significantly across depression severity groups:

$$F = \frac{SSB/(k-1)}{SSW/(N-k)}, \tag{3.17}$$

where SSB is the between-group sum of squares, SSW is the within-group sum of squares, k is the number of groups, and N is the total number of observations.

3.10.2 Post Hoc Analysis

Tukey's HSD test is used for pairwise comparisons:

$$HSD = q(\alpha, k, N-k)\sqrt{\frac{MSW}{n}}. \tag{3.18}$$

3.10.3 Paired t-Test for Model Comparison

Paired t-tests with Benjamini–Hochberg correction are used to compare baseline and entropy-enhanced model performance across cross-validation folds:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{3.19}$$

where $\bar{d}$ is the mean performance difference, s_d is the standard deviation of differences, and n is the number of folds.

3.11 Computational Environment

All experiments are implemented in Python 3.10. The computational environment includes NumPy, pandas, Scikit-learn, XGBoost, TensorFlow, EntropyHub, Antropy, Matplotlib, Seaborn, and SciPy.

3.12 Chapter Summary

This chapter described the complete methodological framework used in the study, beginning with PHQ-9 data collection and preprocessing, followed by entropy-based feature extraction, machine learning classification, unsupervised pattern discovery, threshold optimization, and statistical validation.

Chapter 4

Results and Analysis

4.1 Introduction

This chapter describes the experimental results based on the proposed entropy-enhanced framework for assessing depression severity. The objective of the chapter is to assess whether the implementation of entropy-based and statistical response-pattern features can enhance classification performance beyond traditional PHQ-9 sum-score-based assessment. Besides predictive assessment, this chapter also considers the statistical significance of entropy within depression severity groups, the hidden response subtypes through unsupervised analysis, and the clinical usefulness of entropy as a screening signal.

4.2 Overall Classification Performance

The initial aim of the experimental analysis was to compare the predictive accuracy of the standard PHQ-9-based evaluation with the proposed entropy-enhanced feature representation. The results reveal that the entropy-enhanced feature set was consistently superior to the traditional PHQ-9 sum-score baseline across all tested models.

Table 4.1: Performance Comparison Between Baseline and Entropy-Enhanced Feature Set

Model	Baseline Acc.	Proposed Acc.	Baseline Prec.	Proposed Prec.	Baseline F1	Proposed F1
Random Forest	0.812	0.992	0.805	0.991	0.808	0.992
XGBoost	0.795	0.988	0.791	0.987	0.793	0.988
SVM (RBF)	0.764	0.976	0.758	0.974	0.760	0.975
Neural Network (ANN)	0.782	0.982	0.775	0.980	0.778	0.981
Gradient Boosting	–	0.985	–	0.985	–	0.985
Logistic Regression	0.791	0.975	0.786	0.973	0.788	0.974

The Random Forest classifier produced the strongest performance, increasing accuracy from 0.812 in the baseline configuration to 0.992 with the entropy-enhanced feature set. This improvement indicates that the conventional PHQ-9 total score does not use all the information available in the questionnaire responses. The enhanced feature representation captures response diversity, irregularity, dispersion, and symptom-organization patterns that remain hidden when only the final PHQ-9 score is considered.

4.3 Computational Complexity and Practical Efficiency

Predictive improvement is important, but computational efficiency also counts toward the practical usefulness of an enhanced framework. Considering the fact that the proposed method doubles the number of input features from 9 to 18, it is necessary to consider whether the performance growth is attained at an impractical computational cost.

Table 4.2: Computational Complexity Comparison Using Item-Level Features and Enhanced Feature Set

Classifier	**Baseline (9 Features)**	**EFS (18 Features)**	**Interpretation**
Random Forest	$O(T \times 9 \log n)$	$O(T \times 18 \log n)$	Linear increase, strong performance-cost tradeoff
XGBoost	$O(T \times 9n)$	$O(T \times 18n)$	Parallelizable, scalable in practice
Gradient Boosting	$O(T \times 9n)$	$O(T \times 18n)$	Moderate cost, stable convergence
SVM (RBF)	$O(n^2 \times 9)$	$O(n^2 \times 18)$	Highest computational cost
ANN (MLP)	$O(n \times 9h)$	$O(n \times 18h)$	Linear growth with features and hidden units
Logistic Regression	$O(n \times 9)$	$O(n \times 18)$	Lowest cost, highly efficient

Because PHQ-9 is a short nine-item questionnaire, the enhanced 18-feature representation remains computationally feasible. The additional feature extraction step therefore provides a meaningful improvement in classification performance without making the framework impractical for screening-oriented use.

4.4 Statistical Validation of Entropy Across Severity Levels

To determine whether entropy values differ significantly across depression severity groups, a one-way ANOVA was performed. The analysis showed a statistically significant group effect:

$$F(4, 2034) = 41.73, \quad p < 0.001, \quad \eta^2 = 0.076. \tag{4.1}$$

This finding implies that entropy does not remain constant across levels of depression severity and that the differences are not caused by random error. Notably, the trend of entropy was not linear. Instead of rising steadily with severity, entropy reached its highest level in moderate depression and decreased in severe depression. This trend is an indication of something interesting- in moderate cases of depression, the patient can experience their symptoms shifting and scattered, appearing in unpredictable configurations, whereas in severe cases the patient may find themselves having a much heavier, more homogeneous load to bear, with almost everything suddenly and strongly affected.

4.5 Post Hoc Analysis of Group Differences

The ANOVA test shows us that there is a difference among the groups but it does not tell us exactly where this difference is. So we did a Tukey HSD post hoc analysis to learn more. This analysis helps us figure out which groups are really different from each other. We wanted to know which groups have meaningful differences so we used the Tukey HSD post hoc analysis to find out. The ANOVA test is useful. The Tukey HSD post hoc analysis gives us more details about the differences, between the groups.

Table 4.3: Post Hoc Tukey HSD Results for Entropy Differences

Comparison	**Mean Difference**	**p-adj**	**Significant**
Moderate vs. No Depression	-0.4367	0.000	Yes
Moderately Severe vs. No Depression	-0.4119	0.000	Yes
Moderate vs. Severe	-0.1465	0.000	Yes
Moderately Severe vs. Severe	-0.1216	0.000	Yes
Moderate vs. Moderately Severe	-0.0249	0.134	No

When we look at the results after the fact we can see that the groups with levels of severity are actually different in important ways when it comes to entropy. But there is something to note. The groups that are moderate and moderately severe are not really different, from each other. This means that how complicated the responses are does not always fit into the categories that we usually use to describe severity. This actually makes the main point stronger. The way people answer the PHQ-9 questions tells us something. The PHQ-9 responses carry information that we might miss if we just look at the score.

4.6 Unsupervised Pattern Discovery and Cluster Analysis

More than merely categorizing the responses, based on known degrees of severity, the study went a step further by posing a more open ended question: are there naturally occurring subgroups sweeping op within the data ones that were not difinited by clinical categories to begin with? To investigate this, t-SNE visualization, and K-Means clustering were introduced. What was created was impressive in that there were four different phenotypes that emerged upon the paper, creating the even appearing level of depression that were all on the same level when paper.

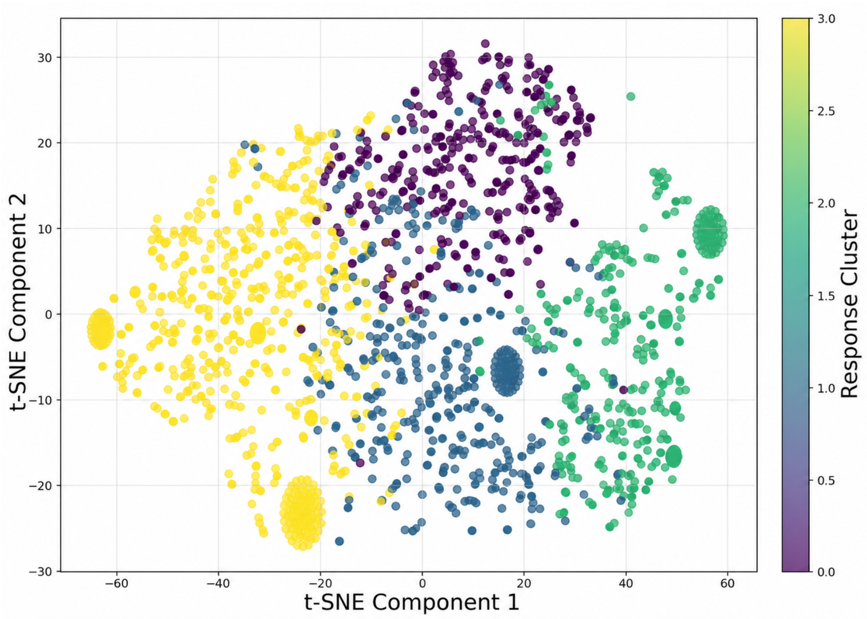

Figure 4.1: t-SNE Visualization of Response Pattern Phenotypes

Table 4.4: Response Pattern Cluster Analysis

Cluster	Depression Score	Shannon Entropy	Sample Entropy	Depression Level
0	16.87 ± 2.83	0.480	0.366	Moderately Severe
1	15.13 ± 2.43	0.377	0.336	Moderately Severe
2	23.27 ± 2.35	0.277	0.414	Severe
3	7.27 ± 3.12	0.304	0.415	Mild

Key Observations from Clustering

- Two different groups of people were found in the category and this shows that people with similar scores do not always have similar experiences with the Patient Health Questionnaire 9.
- The group with the severe symptoms had the highest scores but they also had the lowest variation in their symptoms, which means that they had a consistent and heavy burden of symptoms.

- The group with symptoms had lower scores but their symptoms were all over the place and hard to predict which was a surprise.
- To put it simply the Patient Health Questionnaire 9 scores do not always mean that people have the type of symptoms the Patient Health Questionnaire 9 scores only tell part of the story, about the Patient Health Questionnaire 9.

4.7 Entropy Threshold and Clinical Utility

To see if entropy really works in the world especially for finding severe depression early the study did something called ROC analysis and threshold optimization. The study used something called Youdens J statistics to find the entropy threshold. This helped the study find the level of entropy that can signal when someone has severe depression. The study found an entropy threshold of

$$\tau = 0.439. \tag{4.2}$$

At this point the framework gave us these results:

- Sensitivity = 0.771
- Specificity = 0.793
- AUC = 0.852

Table 4.5: Clinical Performance at the Optimal Entropy Threshold

Metric	**Value**
Sensitivity	0.771
Specificity	0.793
Threshold	0.439
AUC	0.852

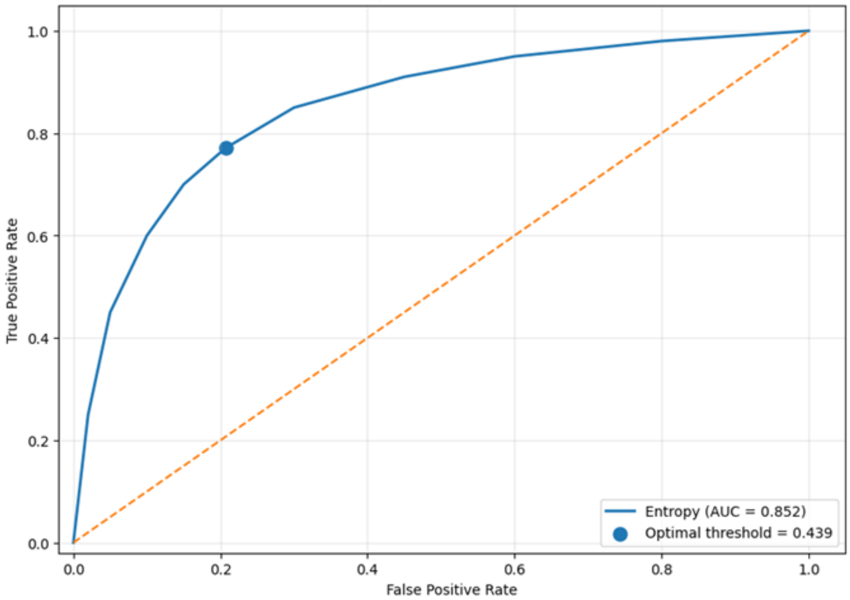

Figure 4.2: ROC Curve: Entropy for Detecting Severe Depression

The corrected ROC result really supports the idea that entropy's not just a statistical oddity it actually helps tell severe depression apart from other levels with an AUC of 0.852. This is a sign that it can screen well and it really supports the idea that entropy could be a useful extra tool in clinics. Its not meant to replace existing tools but to add another layer of understanding that traditional scoring might not catch.

The entropy helps distinguish depression and it could serve as a valuable signal, in clinical settings the entropy is a valuable complementary signal. The threshold of 0.439 should not be interpreted as a standalone diagnostic rule; rather, it provides an additional response-complexity marker that may support PHQ-9 interpretation when used with established clinical screening procedures.

4.8 Discussion of Key Findings

Several major conclusions can be drawn from the results presented in this chapter. First, the entropy-enhanced feature set significantly improves depression severity classification across all evaluated machine learning models, with Random Forest producing the strongest performance. Second, entropy is shown to be statistically meaningful across severity groups, but its relationship with depression is not linear. Third, clustering reveals that people with similar total PHQ-9 scores may still belong to different response-pattern subtypes. Fourth, entropy demonstrates practical utility as a screening indicator through threshold optimization and ROC analysis, with the corrected ROC result showing an AUC of 0.852.

4.9 Chapter Summary

This chapter showed what happened when the proposed entropy-enhanced framework was used to measure the severity of depression. Overall, the results show that response-pattern complexity provides important additional information about depression severity. The classification results, statistical validation, clustering findings, and ROC-based threshold analysis collectively support the view that depression should be understood not only in terms of symptom burden, but also in terms of the internal organization of symptoms.

Chapter 5

Case Studies

5.1 Introduction

This chapter showcases curated case studies from contemporary digital mental health research on depression, aiming to link the proposed framework of this study with practical applications. The objective of this chapter is not solely to summarize previous studies, but to assess how these real-world examples support the broader shift toward data-driven, technology-assisted mental health evaluation and treatment.

5.2 Criteria for Case Study Selection

The case studies were selected based on three criteria. First, the selected studies had to focus directly on depression or depressive symptoms rather than on general mental health alone. Second the studies had to use behavioral or questionnaire-based data in a way that really connects with this studys focus on using computers to analyze things. They had to do more than scratch the surface of computational analysis. They had to really dig in and understand it. Third the selected studies had to be about life and show how digital tools, machine learning or using data to make decisions can actually help with screening, monitoring or supporting people rather than just being ideas that do not really work. The digital tools, machine learning and data driven approaches had to make a difference in the world and the studies had to show this so the studies, on digital tools machine learning and data driven approaches were chosen because they did that.

5.3 Case Study 1: Smartphone-Based Depression Identification Using Minimal App Usage Data

The work that Ahmed and Ahmed did is a good example of how digital mental health research is being used. They made a smartphone system called Mon Majhi that can find out if university students are depressed. What is really cool about this system is that it is very fast. It can look at what apps you used over the seven days in just one second. Then it uses this information to figure out if you’re depressed or not. Ahmed and Ahmed got one hundred university students in Bangladesh to take part in their study. They looked at the students PHQ- responses. Also, at how they really used their smartphones. This helped them get an understanding of the students mental health. The results of the study were very good.

The best model they used which is called gradient boosting was able to correctly identify eighty-two point four percent of the students who were depressed. It was also able to be seventy-five percent of the time and it had an F1-score of seventy-eight point five percent. The study also did something important which is that it tried to be transparent. Ahmed and Ahmed used something called SHAP-based interpretation to explain how their model was making its decisions. This helped them understand which behaviors were actually causing the model to say that a student was depressed. The Mon Majhi system is an example of how digital mental health research can be used to help people.

Why This Case Matters

- It shows that even a little digital data about how people behave can help find depression.
- The study uses the PHQ-9 method, which makes it relevant, to our research.
- Machine learning can really help doctors find depression easily.
- This study agrees with the idea that mental health systems should be easy to use and not too expensive and work well in life.

5.4 Case Study 2: Digital Data-Driven Therapeutic Intervention for Depressive Symptoms

Another good example is, from Fatouros et al,[33] They tested an intervention that used data to help people. This was done through a study where some people got the intervention and others did not. They wanted to see if it could really help with depression and anxiety symptoms in life. This study examined a 16-week decentralized, randomized, waitlist-controlled intervention that integrated wearable sensor data with a mobile application to deliver personalized CBT-based support. The study enrolled 200 adults, with 164 completing the intervention, and used PHQ-9 and GAD-7 measures to track symptom change over time. The results showed that the intervention group experienced significant reductions in depressive symptoms, with a mean change of -5.61 compared with controls, alongside significant anxiety reduction and strong engagement.

Practical Relevance to the Present Study

- Confirms the value of digital, data-driven mental health systems.
- Shows that PHQ-9-based symptom tracking can be integrated into intervention workflows.
- Supports the idea that richer mental health representations may improve personalization.

- Strengthens the argument that computational methods have clinical as well as analytical value.

5.5 Case Study 3: Scalable E-Mental Health Intervention Across Trial and Real-World Settings

A third highly relevant case is the study by Xu et al. [34], which examined a scalable e-mental health intervention for depressive symptoms using both a randomized controlled trial and large real-world datasets. The research included three studies: a randomized trial with approximately 94 wait-list participants and 93 intervention participants in the analyzed comparison, a real-world user study with 11,554 first-time users, and a second large-scale real-world study with 44,018 first-time users. Across these settings, depressive symptoms were measured using PHQ-9, and the intervention combined guided mindfulness and therapeutic writing practices within a structured e-mental health program.

Why This Case Is Strong for the Report

- Demonstrates scalability from controlled trial to large naturalistic settings.
- Uses PHQ-9 repeatedly as a structured outcome measure.
- Connects depression assessment with real-world digital mental health deployment.
- Shows why improving PHQ-9 interpretation can have value in large-scale systems.

5.6 Comparative Discussion of the Case Studies

Taken together, the three case studies reveal several important patterns. First, all of them rely on structured, measurable mental health data rather than only unstructured clinical impressions. Second, each study combines PHQ-9 or related depressive symptom measures with some form of computational or digital system. Third, all three studies show that depression-related information becomes more useful when it is analyzed in richer and more contextual ways—whether through smartphone behavior, wearable-integrated interventions, or scalable app-based programs.

Table 5.1: Comparative Summary of Selected Case Studies

Case Study	Main Focus	Data Source	Role of PHQ-9	Relevance to Present Study
Ahmed and Ahmed [31]	Smartphone-based depression identification	App usage + PHQ-9	Screening label	Shows value of digital behavioral features
Fatouros et al. [33]	Digital therapeutic intervention	Wearable + app + questionnaires	Symptom monitoring	Shows practical role of structured digital assessment
Xu et al. [34]	Scalable e-mental health intervention	App-based intervention + PHQ-9	Outcome measure	Shows scalability of PHQ-9-based digital systems

5.7 Implications of the Case Studies for the Present Research

The selected case studies strengthen the practical relevance of the present report in three ways. These studies are really important because they tell us some truths about the PHQ-9. The PHQ-9 is not a questionnaire that doctors use. It is an useful tool that helps us learn more about mental health and it is used in many real-life situations. The studies also show us that machine learning and data, from smartphones and wearable systems are very helpful. These things are moving the field of health in a very exciting direction. The PHQ-9 and these new technologies are great. The studies also show us that there is still a lot to learn. We need to understand what people are really telling us when they answer questions not just look at their scores. The PHQ-9 is a tool and it can help us do that.

5.8 Chapter Summary

This chapter looked at three examples from recent studies on digital mental health. Each example shows us where this field is going. When we look at all three examples together they give us an idea. Digital mental health is becoming more connected to machine learning the things people do devices that people wear and online help. This shows a change in how we think about and deal with mental health. At the time these examples remind us of something that has not changed. The PHQ-9 is still a trusted tool that many people use to measure health. It is still a part of research even as the technology, around it changes.

Chapter 6

Discussion

6.1 Introduction

This chapter takes a step back from the numbers. Looks at what the findings really mean. The discussion is some big questions. What made the proposed framework work well? What does the weird behavior of entropy tell us about how depressive symptoms show up? How do the different response patterns that we found change our understanding of depression? What does all of this mean, for the future of health screening systems? The discussion explores these questions and more looking at the picture of mental health and the proposed framework and depressive symptoms and the future of mental health screening systems.

6.2 Interpretation of the Improved Classification Performance

One of the interesting things we found out from this research is how much better the classification performance got when we started using entropy-based features. The Random Forest model by itself went from being about 81.2% of the time using just the PHQ-9 total scores to being right about 99.2% of the time when we added entropy and statistical response-pattern features. This is not a big technical deal it also means something important. It shows that the usual way of assessing things by adding up scores, which is useful misses some important information. This information was always there, in the PHQ-9 responses we just had to find a way to uncover it. The PHQ-9 responses had more to tell us and using entropy-based features helped us hear it.

This observation augers quite well with what we said in the report. Not everyone has depression. It has little to do with an individual to another. This implies that the score of the severity of their depression can be shared by two people but their symptoms could be very different. They can also have these symptoms in an order or have more of some symptoms than others. By examining the complexity of the symptoms we are able to obtain an idea of what is being done. This is more helpful than looking at the total score. The overall score fails to reveal the differences and patterns, in the symptoms that people are experiencing. Depression is complex. Does not always follow a simple pattern. Our current look at it assists us to observe this complexity and comprehend it even more.

6.3 Meaning of the Nonlinear Relationship Between Entropy and Depression Severity

One of the surprising things we found in this research is that the relationship between entropy and depression severity is not simple. When we looked at the statistics we saw that entropy does not just go up as depression gets worse. Instead people with depression had the highest levels of entropy. People with severe depression actually had lower levels of entropy. This shows that the connection between depression severity and response complexity is not straightforward. A possible reason for this pattern is that moderate depression may be a state where symptomsre more mixed and not consistent across different areas of a persons life. It is like the experience is still changing and not settled. Severe depression on the hand may bring a kind of consistency, where symptoms are similar. The response pattern is less varied. No less serious. Depression severity and entropy are connected,. In a complex way. Depression severity affects entropy and entropy reflects depression severity. The relationship, between depression severity and entropy is not easy to understand.

6.4 Discussion of Hidden Response Subtypes

The unsupervised clustering results add a lot of value to what the data's showing. The four distinct response patterns that emerged suggest that PHQ-9 responses have a hidden subgroup structure. This structure goes beyond the clinical severity categories. What stands out is that within the severe range two subtypes appeared. These individuals had identical PHQ-9 scores but their entropy profiles were different. This shows that regular severity labels might be grouping people together. Their internal symptom organization is actually quite different from one another. A simple score wouldn't reveal this difference on its own. The PHQ-9 response patterns are important. They help us understand the data better. The clustering results give us insights into PHQ-9 responses. The subgroups, in PHQ-9 responses are meaningful.

6.5 Relevance to Existing Literature

The results of this study match what other researchers have been saying. They have been questioning whether the usual way of scoring tests really shows the picture [14]. Other studies have found that the total score on the PHQ-9 test can hide differences in how symptoms appear. They also found that looking at each question on the test can reveal patterns that scoring the test cannot. This study takes it a step further by showing that looking at each question can improve how well we can tell how severe someones symptoms are. It also shows that this approach can find patterns and groups that the total score cannot.

This study is also related to recent work in behavioral science and psychiatry. These fields are finding that mental health data does not always follow a straight line [12]. This study is part of a movement in digital mental health and research, on digital biomarkers. The field is moving towards using data and better ways of understanding mental health conditions. The PHQ-9 test and mental health research are becoming more complex and detailed. The study of health conditions and the PHQ-9 test are getting more attention and more research is being done to understand them better.

6.6 Practical and Clinical Implications

What makes this study really important is that it does not need to get rid of a tool that doctors already use and like. The new framework is not meant to replace the PHQ-9 it is meant to work with the PHQ-9 and add information about how people answer questions. This is a strong idea: if we look at the answers to the questions in a smarter way not just the total score then we can get more information from the tests without making it harder for people to fill them out.

The idea of an entropy threshold is also very useful. The best cutoff point of 0.439, for finding depression means we can use entropy to help identify the kinds of symptoms that are often seen in severe cases. The corrected analysis supports this idea with a score of 0.852 which shows that this approach is working. This does not mean entropy should be used alone to diagnose depression. It shows that a measure of complexity can be used in a way that is meaningful and easy to understand in a clinical setting. The PHQ-9 and entropy can work together to help doctors understand the PHQ-9 results better. The study of entropy and the PHQ-9 is important because it can help us use the PHQ-9 in a way.

6.7 Methodological Implications

The study is indeed a good one since it unites various fields of research. The PHQ-9 is a base, for this research. Next there is entropy measures. Assistance (Please, help). Enable us to perceive the complex aspects that people are talking about. Machine learning picks up on this information and determines whether it can assist us in making predictions. Clustering assists us in revealing the clusters that it hides underneath the surface. Statistical tests ensure that what we discover is no less than real and is not an accident. The PHQ-9 and entropy measures, machine learning and clustering, and statistical tests are all collaborative. All these sections are essential. They are all combined in order to make something really useful. The PHQ-9 and machine learning and clustering and statistical tests all form a part of this research.

6.8 Limitations in Interpreting the Discussion

While the conclusions carry real weight, it is important to view them through the lens of the study is design and the context in which they were reached. The information is based on a relatively homogeneous sample of students, and therefore the trends may be influenced by the specifics of the sample. In addition, the results are obtained through internal cross-validation and are not externally validated on a different clinical dataset. Therefore, the unusually high classification performance should be viewed as evidence of strong internal validity but should not be interpreted as evidence that the findings apply to the entire population.

6.9 Chapter Summary

This chapter explained the key results of the research in relation to depression assessment, response-pattern complexity, and the broader literature. Overall, the discussion confirms the main conclusion that depression severity should be viewed not only through the lens of symptom burden, but also through the lens of symptom arrangement.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

This paper aimed to investigate whether depression severity assessment could be enhanced by moving beyond the traditional PHQ-9 sum-score method and by considering the hidden complexity of item level response patterns. The results show that this goal was really accomplished. The proposed framework brought together metrics that are based on entropy, statistical descriptors and machine learning approaches. This framework gave a clearer and more detailed picture of PHQ-9 responses. It did a job of classifying results than the traditional score-based approach.

The study made one thing very clear: entropy is not a complicated math idea that is added to the analysis. It is actually an useful tool for evaluating mental health. The statistical analysis showed that entropy changes a lot across levels of depression severity. The post hoc results and clustering analysis showed that this relationship is not simple. Entropy was highest at depression and then decreased at severe depression. This finding goes against the idea that more symptoms automatically means mixed up or varied responses. In fact severe depression seems to make symptoms more consistent and uniform than all, over the place. PHQ-9 responses were more predictable when depression was severe.

Another important thing to note is how the study combines its methods. The framework uses a mix of tools including an used questionnaire, math-based measurements, predictive modeling, grouping and statistical checks. These are all brought together into one understandable process. Importantly this doesn't replace the PHQ-9. Instead the study shows what can be achieved when computer analysis is used to get insight from what the PHQ-9 already provides. The study makes an argument, for using computational analysis to unlock a deeper layer of meaning from the PHQ-9.

7.2 Future Work

Though the results of this paper are promising they also create an avenue to a set of constructive directions that can be developed in future studies. Outside validation of the framework comparing them to independent datasets of clinically diverse patients constitutes the most immediate need. As the current work is based on the responses of the university students, subsequent studies must go beyond the same boundary and explore the extent to which the methodology can be held steady among the larger groups of the population (such as various age groups, clinical patients, diverse communities, etc.). The second significant direction is to introduce a longitudinal aspect to it. This investigation represents the responses of the PHQ-9 at only one point in time but it is rarely the case that a depression remains stationary at any single point in time it moves, intensifies and in some cases lifts up over a period of time. Further studies could examine the dynamics of changes in the value of entropy over repeated evaluations and whether dynamic changes in the complexity of responses can be a predictive indicator of advancement, improvement, or the risk of relapse. The potential of the expansion to multimodal data sources also has a real promise. The existing paradigm operates only under the questionnaire responses but the digital mental-health is expanding rapidly to incorporate passive-sensing, behavioral-logs, physiological-signals and mobile interaction-data [5, 8, 9]. The combination of entropy based PHQ-9 features and these richer inputs may open the way to the fairly more robust, and at the same time more sensitive to the full picture of a person is experience. Regarding the methodological front future work may as well venture into more machine learning and explainability approaches that may help us to gain a deeper insight into how and why the framework functions in the way that it does. And lastly, perhaps one of the most valuable opportunities in the future is to understand how an entropy-based analysis of response might one day be applied to offer truly personalized mental health care to an individual past merely responding to an aggregate trend and instead begins to engage with the individual.

Bibliography

[1] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[2] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[3] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039–H2049, 2000.

[4] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical Review Letters*, vol. 89, no. 6, p. 068102, 2002.

[5] T. R. Insel, "Digital phenotyping: Technology for a new science of behavior," *World Psychiatry*, vol. 16, no. 2, pp. 121–122, 2017.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[7] C. Bandt and B. Pompe, "Permutation entropy: A natural complexity measure for time series," *Physical Review Letters*, vol. 88, no. 17, p. 174102, 2002.

[8] N. C. Jacobson et al., "Digital biomarkers of passive optical sensor data in predicting depression," *npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[9] J. Torous et al., "Digital biomarkers for mental health: A vision for the future," *Nature Mental Health*, vol. 1, no. 2, pp. 12–24, 2023.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[12] Y. Cho et al., "Nonlinear dynamics of depressive disorders: A review of entropy applications," *Entropy*, vol. 22, no. 7, p. 748, 2020.

[13] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.

[14] B. Shields et al., "Beyond sum-scores: New metrics for PHQ-9," *Psychometrika*, vol. 85, no. 4, pp. 912–930, 2020.

[15] M. Ghassemi et al., "Opportunities in machine learning for healthcare," *Machine Learning*, vol. 109, no. 1, pp. 1–5, 2020.

[16] Y. Sun, Z. Fu, Q. Bo, Z. Mao, X. Ma, et al., "The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital," *BMC Psychiatry*, vol. 20, art. 474, 2020.

[17] Y. Sun, Z. Kong, Y. Song, J. Liu, et al., "The validity and reliability of the PHQ-9 on screening of depression in neurology: A cross-sectional study," *BMC Psychiatry*, vol. 22, art. 98, 2022.

[18] E. Beswick et al., "The Patient Health Questionnaire (PHQ-9) as a tool to screen for depression in people with multiple sclerosis: A cross-sectional validation study," *BMC Psychology*, vol. 10, art. 281, 2022.

[19] M. Brattmyr et al., "Factor structure, measurement invariance, and concurrent validity of the Patient Health Questionnaire-9 and the Generalized Anxiety Disorder scale-7 in a Norwegian psychiatric outpatient sample," *BMC Psychiatry*, vol. 22, art. 461, 2022.

[20] X. Gao and Z. Liu, "Analyzing the psychometric properties of the PHQ-9 using item response theory in a Chinese adolescent population," *Annals of General Psychiatry*, vol. 23, art. 7, 2024.

[21] S. Tomitaka et al., "Distributional patterns of item responses and total scores on the PHQ-9 in the general population: Data from the National Health and Nutrition Examination Survey," *BMC Psychiatry*, vol. 18, art. 108, 2018.

[22] F. Cosci et al., "Patient Health Questionnaire-9: A clinimetric analysis," 2024.

[23] K. H. Kristofersdottir et al., "Disclosing depressive symptoms: Perceived sensitivity of PHQ-9 items in a general population sample," *BMC Psychology*, vol. 14, art. 294, 2026.

[24] O. F. Ayilara et al., "Sensitivity of the Patient Health Questionnaire (PHQ-9) items to change in three clinical trials in low- and middle-income countries," *Journal of Affective Disorders*, vol. 393, pt. A, art. 120226, 2026.

[25] C. Nunez et al., "Understanding symptom profiles of depression with the PHQ-9 in a community sample using network analysis," *European Psychiatry*, vol. 67, no. 1, e50, 2024.

[26] "The same but different too: Depression profiles in young adults without a history of psychiatric treatment identified using Bayesian and partial correlation networks," *Journal of Psychiatric Research*, 2024.

[27] N. C. Jacobson and S. Wilhelm, "Digital biomarkers of mood disorders and symptom change," *npj Digital Medicine*, vol. 2, art. 3, 2019.

[28] V. De Angel et al., "Digital health tools for the passive monitoring of depression: A systematic review of methods," *npj Digital Medicine*, vol. 5, art. 3, 2022.

[29] S. Akre et al., "Advancing digital sensing in mental health research," *npj Digital Medicine*, vol. 7, art. 362, 2024.

[30] S. Mulinari, "Aligning digital biomarker definitions in psychiatry with the National Institute of Mental Health Research Domain Criteria framework," *NPP—Digital Psychiatry and Neuroscience*, vol. 2, art. 15, 2024.

[31] M. S. Ahmed and N. Ahmed, "A fast and minimal system to identify depression using smartphones: Explainable machine learning–based approach," *JMIR Formative Research*, vol. 7, e28848, 2023.

[32] P. Alves et al., "A machine learning model using clinical notes to estimate PHQ-9 symptom severity scores in depressed patients," *Journal of Affective Disorders*, vol. 376, pp. 216–224, 2025.

[33] P. Fatouros et al., "Randomized controlled study of a digital data-driven intervention for depressive and generalized anxiety symptoms," *npj Digital Medicine*, vol. 8, art. 113, 2025.

[34] J. Xu et al., "A scalable mental health intervention for depressive symptoms: Evidence from a randomized controlled trial and large-scale real-world studies," *npj Digital Medicine*, vol. 8, art. 491, 2025.

[35] L. Wang et al., "AI-assisted multi-modal information for the screening of depression: A systematic review and meta-analysis," *npj Digital Medicine*, vol. 8, art. 523, 2025.

[36] A. Trelles et al., "Systematic review and meta-analysis of explainable machine learning models for clinical depression detection," *Behavioral Sciences*, vol. 15, no. 11, art. 1476, 2025.

[37] W. Lu et al., "Diagnostic accuracy of traditional and deep learning methods for detecting depression based on speech features: A systematic review and meta-analysis," *BMC Psychiatry*, vol. 25, art. 1190, 2025.

[38] B. Grimm et al., "Multimodal machine learning for video-based single-question mental health assessment," *npj Digital Medicine*, vol. 9, art. 65, 2026.

[39] R. Halabi et al., "A systematic exploration of digital biomarkers for the detection of depressive episodes in bipolar disorder," *npj Mental Health Research*, 2026.

[40] A. Ikaheimonen et al., "Variability in self-reported depression symptomology and associated behavioral markers in digital phenotyping," *Scientific Reports*, 2026.