A Cross-Modal Attention-Based Multimodal Deep Learning Framework for Early Prediction of Adolescent Mental Health Disorder

by

Md. Taosif
22301178
Ummay Maimona Chaman
22301719
Nazifa Anjum Prova
22301675
Sidrat Moon Taher
22301747

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
June 2025.

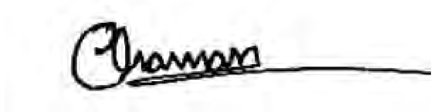
Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

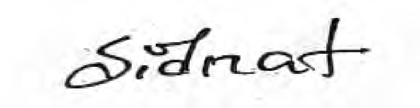
Student's Full Name & Signature:

Md. Taosif
22301178



Ummay Maimona Chaman
22301719

Nazifa Anjum Prova
22301675



Sidrat Moon Taher
22301747

Approval

The thesis/project titled "A Cross-Modal Attention-Based Multimodal Deep Learning Framework for Early Prediction of Adolescent Mental Health Disorder" submitted by

1. Md. Taosif (22301178)
2. Ummay Maimona Chaman (22301719)
3. Nazifa Anjum Prova (22301675)
4. Sidrat Moon Taher (22301747)

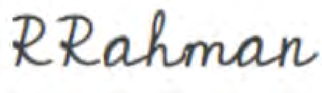
of Spring, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 19, 2025.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Mr. Rafeed Rahman
Senior Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Ethics Statement

This study utilized data from the Queensland Twin Adolescent Brain (QTAB) cohort, adhering strictly to sanctioned ethical standards and data governance protocols. Because adolescent mental health data is sensitive, all analyses were done in a secure, controlled computing environment that QTAB approved. No raw data was duplicated, shared, or accessible outside of the permitted infrastructure. Data utilization was strictly confined to the parameters of this study and sanctioned solely by regulatory consent. All techniques complied with institutional and international ethical guidelines regulating research involving human volunteers and medical artificial intelligence.

Abstract

Adolescent mental health conditions are frequently underdiagnosed due to the limitations of single diagnostic techniques and the reduced efficiency of procedures that fail to combine biological, psychological, and social aspects. We introduce a cross-modal attention-based multimodal deep learning framework that uses structural brain imaging and phenotypic data to identify anxiety disorders in the early stages. The proposed method incorporates the QTAB dataset to concurrently model T1-weighted magnetic resonance imaging (T1w MRI), behavioral questionnaire responses, and demographic variables, revealing related patterns of risk across modalities. The framework consists of three independently trained modality-specific encoders: a 3D convolutional neural network for structural MRI that generates 768-dimensional neuroanatomical representations, and two self-attention enhanced prototypical learning modules for behavioral and demographic data that generate discriminative metric embeddings of 32 and 64 dimensions, respectively. Cross-modal integration happens at the decision level, allowing for more reliable fusion in sample sizes that are limited. The experimental evaluation shows that the proposed framework obtains an AUC of 0.8935 for predicting anxiety disorders, representing a 15.1% improvement over the most robust unimodal baseline (questionnaire: AUC = 0.7766), with 85.7% sensitivity and 87.3% specificity. The best fusion weights were 23% MRI, 63% questionnaire, and 14% demographics, demonstrating the questionnaire data's higher predictive signal. Results demonstrate that optimized weighted late fusion of well-calibrated modality predictions surpasses intricate learnt fusion techniques, underscoring the significance of proper weight optimization and calibration in small-sample multimodal psychiatric modelling. The results illustrate the efficacy of multimodal late fusion in the early detection of anxiety risk in adolescents and offer a clinically significant approach for scalable mental health screening systems.

Keywords: Prototypical Network, Few-shot, Neuroimaging, Twin-split, Multimodal Fusion, Cross-modal, 3D Convolutional Neural Network

Dedication (Optional)

A dedication is the expression of friendly connection or thanks by the author towards another person. It can occupy one or multiple lines depending on its importance. You can remove this page if you want.

Acknowledgement

We convey heartfelt gratitude to our thesis supervisor, Dr. Md. Golam Rabiul Alam, Professor in the Department of Computer Science and Engineering at BRAC University, for his unwavering support, insightful guidance, and constructive feedback during our research journey. His knowledge and support played a crucial role in the development of this work. We express our gratitude to our co-supervisor, Mr. Rafeed Rahman, Senior Lecturer in the Department of Computer Science and Engineering at BRAC University. His technical support, insightful suggestions, and unwavering cooperation were instrumental in the successful completion of this thesis. We want to express our gratitude to the Department of Computer Science and Engineering at BRAC University for offering the essential academic environment and resources that facilitated this research. We extend our gratitude to the Queensland Twin Adolescent Brain (QTAB) cohort for granting access to the dataset. All analyses were performed in strict compliance with approved ethical standards and data governance protocols, ensuring a secure and controlled computing environment. We appreciate the support and insightful discussions provided by our faculty members, friends, and peers.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

3T	3 Tesla
AUC	Area Under the Curve
BRS	Brief Resilience Scale
CNN	Convolutional Neural Network
CRSQ	Children's Response Style Questionnaire
DZ	Dizygotic
F1	F1-Score
GELU	Gaussian Error Linear Unit
MNI	Montreal Neurological Institute
MRI	Magnetic Resonance Imaging
MZ	Monozygotic
NIMH	National Institute of Mental Health
PSS	Perceived Stress Scale
QTAB	Queensland Twin Adolescent Brain
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
SCAS	Spence Children Anxiety Scale
SDQ	Strengths and Difficulties Questionnaire
SMFQ	Short Mood and Feelings Questionnaire
SPHERE	Somatic and Psychological Health Report
T1w	T1-weighted

Chapter 1

Introduction

1.1 Background

Mental health in adolescents becomes a critical problem area in the scope of global public health as it has been reported, approximately one in every five adolescents has a psychiatric disorder such as anxiety and depression [56]. Adolescent mental illnesses stand out as a serious issue of concern to the general population, since most of the signs originate in childhood and adolescence, and may develop in adult life provided that they are not identified or treated. The U.S. National Institute of Mental Health (NIMH) states that as a component of general health, mental health is critical to children and adolescents, and as the causes of many mental disorders in the adult population were manifested in childhood or adolescence, but were not properly resolved at that stage, it is essential to detect and treat them at early stages [31]. Early treatment of emerging mental health problems, such as anxiety, can prevent more severe and long-lasting difficulties later in life [34].

During adolescence, anxiety disorders may be significantly problematic in terms of how one thinks, behaves, and functions in everyday life. These disorders are linked to poor performance in school, social interaction as well as the general quality of life and usually accompany other psychiatric disorders, including depression and behavioral disorders [28]. To illustrate the point, the national survey data indicate that the proportion of all adolescents, who do experience the symptoms of anxiety, is rather high, and the mental health conditions are still prevalent among adolescents, which means that the screening methods and the models of early prediction should be improved [20], [22]. The example of multimodal deep learning models with co-analysis of neural images, genetic and behavioural data has been shown to be more beneficial than unimodal deep learning models [50]. It was also more recently demonstrated that with a trimodal neural network predicting a correct classification of 8 schizophrenia can be realized at 88 percent with a combination of data of sMRI, fMRI, and genomics data [13]. Kanyal et al. (2024) presented a detailed model that looked into the integration of sMRI, fMRI and SNP markers and achieved a rate of an estimated 79 percent, although the discussion that accompanied the results identified important biological variables [15].

Queensland Twin Adolescent Brain (QTAB) research offers a useful and genetically informative longitudinal dataset to determine predictors of mental illnesses in adolescence at a young age [61]. The attraction to twin studies is in the fact that they enable the partial disentangling of genetic and environmental effects on mental health outcome [29], [60] but also have special issues to do with genetic correlation among the subjects [61]. Scales Spence Children Anxiety Scale (SCAS) is an effectively validated 44-item measure of youth anxiety in various symptom domains and an overall score of 30 or higher is a clinical cut-off point setting anxiety in adolescents populations [1]. It is possible that by incorporating behavioral, neuroimaging, and demographic data into a predictive model, one can discover people at high risk prior to the emergence of clinical anxiety to make proactive and preventive intervention decisions. These technologies together offer a great platform to come up with intelligent, responsive and clinically relevant systems, besides predictive systems.

1.2 Rationale of the Study or Motivation

The issue of adolescent mental health disorders, especially anxiety and depression, constitutes an increasing problem in the domain of public health, and it is related to the early onset and chronicity of the disorder, as well as a significant effect on academic, social, and neurodevelopment outcomes [31], [56]. Conventional single-modality, or diagnostic practices, including self-report questionnaires, or clinical interview, are inadequate in their lack of capability to be capable of the multidimensional expressions of these disorders. They tend to be cross-sectional, subjective and incapable of explaining the complicated combination of risk factors (biological, psychological and environmental) [43], [44]. The currently available studies on mental health prediction among adolescents also have various challenges. The majority of methods are mostly single-mode, those that consider either behavioral, textual or neuroimaging, isolated data, limiting predictive validity and immature to reflect the emergent interplay of risk factors [33]. Most models are also fitted to cross-sectional data or adult populations, which restricts their applicability to groups of adolescents [33]. The dependency based on families, or twins, is also frequently overlooked, where inflated performance estimates can be obtained due to genetic confounding [29], [60], [61].
Furthermore, they are not often interpretable, and frequently, they can be poorly interpretable to clinical reasoning because deep learning models are often opaque, which restricts their interpretable potential as well as applicability in practice [20], [47]. Recent developments in multimodal deep learning can offer an opportunity to overcome these shortcomings; behavioral questionnaires, neuroimaging features, and demographic or genetic data are all promising methods to overcome these limitations. These methods may be able to identify latent patterns that show the early risk, enhance the accuracy of prediction and an early intervention to the high-risk adolescents who might be vulnerable to the development of anxiety disorders. The motivation behind this study has been to devise a cross-modal predictive framework, twins aware and leakage Free so as to amalgamate these complementary modalities. The study will determine adolescents at risk of anxiety before clinical onset by using longitudinal data, provided by Queensland Twin Adolescent Brain (QTAB) cohort [58] so that preventive strategies would be applied and the clinical utility of the predictive models should be improved.

1.3 Problem Statement

While much has been done in developing predictive models that can predict mental disorders in adolescents, early detection of mental disorders in this age bracket remains very difficult. Most of the existing methods are limited to one type of data, for example, behavioral features, neuroimaging, or textual input alone; all these limit their capabilities of characterization and representation of multifaceted and complex nature of mental health risk. Typically, many datasets are small and imbalanced, which reduces the integrity of models and raises the chances of overfitting, and familial or twin-related dependencies are often omitted, which provides unrealistic estimates of performance and increases the false uniformity of performance. Additionally, despite deep learning models providing high predictive performance, they are highly inapplicable in clinical settings due to complexity and no interpretability characteristic. The traditional approaches are also overly dependent on cross-sectional testing that only gives examples of actual symptoms, but not a longitudinal study and the risks that may develop in the early life. The disadvantages emphasized the necessity of implementing a twin-conscious, multimodal, predictive model, which can combine neuroimaging, behavioral and phenotypic data that will allow, not only to identify adolescents, who are at risk of developing any type of anxiety disorder, but also do it in a timely manner and be explained and interpreted safely.This study investigates, "Does incorporating twin information (monozygotic and dizygotic pairs) with T1-weighted MRI, anxiety questionnaires, and phenotypic data improve the early prediction of anxiety in adolescents from baseline to the first follow-up, compared to non-twin datasets such as ABCD?"

1.4 Research Objectives

The goal of this thesis is to create a multimodal deep learning framework that is aware of twins and can predict anxiety in teens early on and over time. This method combines neuroimaging, behavioral questionnaires, and phenotypic data to make predictions more accurate, clear, and useful in a clinical setting.
This study aims to achieve the following goals:

- To create a prediction framework that combines T1-weighted MRI data, answers from anxiety questionnaires, phenotypic information to find teens who are likely to have anxiety early on.
- To build a robust MRI modality based forecasting module and feature extraction strategy that employs pretrained 3D convolutional neural networks to identify neuroanatomical patterns linked to anxiety risks.
- To offer a prototype-driven behavioral modeling methodology for the examination of anxiety questionnaires, enabling the accurate detection of clinically significant behavioral patterns.
- To provide a module for phenotypic representation that simulates basic clinical aspects and genetic data, like twin zygosity.

- To use decision-level and cross-modal fusion methods to combine information from different sources in a useful way, and to look at how each source and their combinations affect prediction accuracy.

This thesis tries to solve important problems in predicting the mental health of teenagers by using a twin-aware, multimodal, and longitudinal approach. This study aims to make it easier to find anxiety in teens early on and take preventative steps right away by combining a lot of different data sources. It also stresses the importance of making the data easy to understand and ethical concerns.

1.5 Scope and Challenges

This study investigates teens aged 12 to 18 from the Queensland Twin Adolescent Brain cohort. It utilizes a unique longitudinal twin dataset to distinguish the impacts of heredity and environment on mental health. This study employed a multimodal approach, using standardized anxiety questionnaires for behavioral assessment, structural MRI neuroimaging data, and demographic and genetic variables to clarify the complex factors influencing teenage anxiety. Prototype-based networks offer clinically interpretable veiled representations of vulnerability, clarifying underlying risk profiles that transcend basic prediction accuracy. We use twin-aware assessment methods to make sure that performance indicators really represent how effectively the complete system is operating and to stop co-twin data from leaking. This approach differs from only examining symptoms at a single moment. It doesn't try to guess when anxiety will start. This highlights how vital it is to take actions to stop issues before they happen and how easy it is to make quick decisions regarding the therapy.

This research examines diverse methodological and practical obstacles associated with multimodal, twin-based forecasts of teenage mental health. The small size of the QTAB cohort makes it hard to get enough data, which raises the danger of overfitting and lowers the statistical power needed to find minor variations in anxiety cases. Training models is harder when datasets are of poor quality or change too much, which makes them less reliable. There are problems when trying to combine several forms of data, such behavioral surveys, structural MRI, and phenotypic or genetic information. This is because the scales, dimensions, and noise characteristics are all different. It is important to find a balance between how easy a model is to use and how complicated deep learning is so that the findings are clinically relevant and useful. It is very important to carefully plan how to evaluate twins so that data does not get out of co-twins in twin groups that are affected by genetic bias. In a clinical environment, it is essential to balance sensitivity and specificity to accurately diagnose at-risk youth and provide them with timely early detection treatment.

1.6 Thesis Organization

This thesis consists of six chapters, each focusing on one particular aspect of the research concerning the early prediction of adolescent mental health disorders through a cross-modal attention-based multimodal deep learning framework.

- **Chapter 1: Introduction** provides an overview of adolescent mental health disorders and discusses the importance of early prediction through advanced machine learning techniques. The research problem is identified, the study's main objectives are specified, and the scope, obstacles, and impact of the planned effort are examined.

- **Chapter 2: Literature Review** The Literature Review offers an extensive examination of current studies regarding the prediction of teenage mental health. The conversation includes classical and deep learning approaches, multimodal along with neuroimaging methods, few-shot learning methodologies, and twin-based longitudinal studies. The chapter ends by pointing out important gaps in the study and calling for the use of the proposed methodological procedure.

- **Chapter 3** covers the system requirements and implementation details of the proposed framework, with a focus on its impacts and limits. The conversation covers things like hardware and software installations, ways to measure success, statistical analysis methods, effects on society and the environment, ethical issues, following standards, and checking the project's risk and financial health.

- **Chapter 4: Methodology** The methodology provides a full explanation of the suggested cross-modal multimodal deep learning framework. This paper provides an overview of the dataset employed, delineates the target variable, elucidates the twin-aware data splitting methodology, specifies the data processing modules for every modality (MRI, questionnaire, and static features), discusses the prototypical network architectures, clarifies the design of the loss function, and introduces the multimodal fusion strategy.

- **Chapter 5: Result Analysis** The analysis of the results examines how well the proposed framework functions and talks about the experiment's results. This study examines the effectiveness of both single-modality and multimodal fusion, evaluates statistical significance, conducts ablation tests, and compares the results with current leading methodologies, resulting in a thorough analysis of the findings.

- **Chapter 6: Conclusion** The conclusion points out the most important results and contributions made by this research. The suggested paradigm accurately forecasts adolescent mental health illnesses at an initial stage and offers suggestions for subsequent research opportunities.

Chapter 2

Literature Review

2.1 State of Art in Adolescent Mental Health Prediction

Recent studies have increasingly utilized deep learning along with machine learning methods to enhance the early detection of mental health issues in adolescents. These methods go beyond standard clinical evaluations by incorporating a data-driven examination of psychological, behavioral, and neurobiological characteristics. Recent improvements in multimodal computation and representation learning have made it easier for us to spot patterns of danger, which has led to early screening and the creation of tailored therapeutic methods.

2.1.1 Adolescent Mental Health Disorders and Early Prediction

Mental health disorders, especially anxiety among adolescents, pose significant challenges to global public health. Their early onset, chronic nature, and detrimental effects on academic performance, social interactions, and neurodevelopment are concerning. Early adolescence is a crucial developmental phase characterized by the interplay of psychological vulnerability, brain structure, and environmental influences. Consequently, it is essential to screen high-risk patients early, before clinical progression, to allow for the implementation of timely and effective safeguards.

Conventional methods of screening mainly use symptom-based tests that are conducted once the disorder has occurred, and thus they are not useful in the prevention of the disorder at an early stage. In turn, recent studies have paid more and more attention to the field of early risk prediction based on data-driven techniques exploiting psychological, behavioral and contextual features. Longitudinal surveys of adolescents have been used to find that machine learning models could have moderate predictive accuracy of future mental health trajectories (AUC 0.80-0.87) [36], [57]. Nonetheless, these models are not always as explainable, population-biased, and based on self-reported measures as limited.

Nevertheless, in most cases of adolescent-oriented prediction research, the approach is predominantly unimodal and does not include any objective biological data including neuroimaging. Additionally, the genetic confounding (which is particularly important with family- or twin-based cohorts) is often not estimated; there are concerns about overestimating performance and decreased external validity. These deficiencies highlight the importance of having objective, multimodal, and developmentally based prediction models that clearly indicate the involvement of genetic and environmental components in modeling mental health risks in adolescents.

2.1.2 Machine Learning and Deep Learning in Mental Health Prediction

Machine learning (ML) and deep learning (DL) approaches have shown promising results in predicting mental health outcomes utilizing data in a variety of formats. Text-based data has been processed using techniques such as TF-IDF encoding and Word2Vec embedding alongside classic models such as Support Vector Machine (SVM) and Random Forest. Such approaches have yielded AUC values ranging from 0.84 to 0.89 for data sets such as the Reddit dataset and the CLPsych dataset, demonstrating the potential for demonstrating the benefit of early screening based on language [43], [44]. Recent techniques using transformer-based models as encoders have improved the potential for the comprehensiveness of data interpretation with an AUC as much as 0.92 but remain constrained to language-based data [28].
Beyond textual data, speech- and audio-based models integrating acoustic and paralinguistic features have demonstrated reliable predictive performance, with AUCs ranging from 0.81 to 0.90 across DAIC-WOZ and clinical voice datasets [35]. The use of large language models combined with the inclusion of speaking timing features reinforced the accuracy of classifications (∼91%) while substantially decreasing false positives, again reinforcing the potential utility of the 'speech' phenotype in early detection [19], [59]. Passive sensing approaches relying on smartphone/wearable device data, e.g., activity, sleep, GPS, have achieved up to 0.89 AUC values along with recall rates close to 0.85, allowing continuous unobtrusive surveillance of adolescent mental health states [42], [49].
The advances notwithstanding, the majority of current ML and DL methods are still either unimodal or weakly cross-sectional/mechanistically based. Classical ML methods, while interpretable and computationally efficient, are constrained by handcrafted features and limited capacity to model complex nonlinear relationships, particularly under heterogeneous adolescent datasets [20], [43], [44]. As a result, there is a growing need for multimodal along with biologically informed predictive frameworks that integrate behavioral, neuroimaging, and genetic data to improvise early risk detection and clinical utility.

2.1.2.1 Textual and Social Media–Based Mental Health Detection

The interest in text-based mental health detection utilizing social media channels like Twitter and Reddit has witnessed growing popularity in recent times, mainly because such channels have easily accessible data sufficient enough to identify signals at an early stage. Models based on techniques like TF-IDF, Word2Vec, and transformer models like BERT have reported values between 0.88 and 0.92 in their AUC values, along with accuracy rates closer to 89% [28], [43]. The development of auxiliary approaches like the use of time-related signals increases the accuracy of their use, like in some studies based on BiLSTM approaches that reported accuracy closer to 74.55% in their validation phase while using large-scale data collected via social media channels like Reddit [45].

2.1.2.2 Speech, Audio-Visual, and Behavioral Signal Analysis

The assessment of prosody in human speech, micro-expressions on faces, and behavioral dynamics act as non-intrusive digital phenotypes in mental health assessment. CNN-LSTM fusion-based CNNs that incorporate unimodal speech prosody data along with facial micro-expression data have demonstrated higher accuracy than unimodal approaches, yielding balanced F1-scores of 0.88-0.90 [35], [51]. More recent approaches adopting the use of Retrieve Augmented Generation-based LLMs that incorporate features of timing in unimodal verbal data have reported accuracy rates of over 91%, along with significant reduction rates in false positive outcomes [19]. These approaches still largely remain conditional upon data recording conditions, along with unimodal-verbal linguistic limitations.

2.1.2.3 Deep Learning for Multimodal Mental Health

Recent breakthroughs in DL have greatly contributed to better accuracy in mental health prediction through feature learning directly from raw data. CNN-LSTM architectures have been used in many studies to combine speech signals, facial expressiveness, and text data, obtaining accuracy in the range of 86-91%, along with the highest values of AUROC equal to 0.93 in benchmarks like DAIC-WOZ and AVEC [25], [40], [51]. The utilization of transformer architectures has obtained better performance through higher values of AUROC, even higher than 0.91, in the evaluation of depression estimation studies [47]. However, research has mostly been based on restricted interviews, is still sensitive to incomplete data sets, and has poor ability in addressing adolescent cohorts.

2.2 Mental Disorder Prediction Using Neuroimaging-Based Deep Learning

Neuroimaging provides an important biological understanding of brain structure and function, enabling objective markers to be used to understand mental health vulnerability. Deep learning techniques, especially CNNs, have been adopted to analyze sMRI, fMRI, DTI, and EEG modalities to explore hidden features for psychiatric predictions. A longitudinal study among adolescent cohorts has shown encouraging predictive power for depression prediction, as provided in the study based on the ABCD study design, where an average AUROC score ranged from 0.62 to 0.72 and depression prediction was based on fMRI and structural modalities [54]. Furthermore, improved performance (88%) was available through incorporating genetic information provided using both structural and functional imaging modalities for predicting schizophrenia [13].

Nevertheless, neuroimaging-based approaches are likely to have high specificity and low sensitivity. Such potential pitfalls are not only seen during early risk prediction but are also apparent in the limited sample size or variations between sites and contexts. This further necessitates achieving higher sensitivity and generalizability through neuroimaging-based approaches along with behavioral variables and demographic or genetic features.

2.2.1 Early Prediction of Adolescent Mental Health Disorders Using Multimodal Deep Learning

Previous investigations examined the early identification of mental health disorders in adolescents through various multimodal and behavioral signals, yielding beneficial but not entirely comprehensive findings. Multimodal deep learning frameworks that integrate audio, video, text, facial expressions, voice, and physiological features have demonstrated a UAR of 0.709 for bipolar disorder and an F1 score of up to 0.917 for depression, with accuracy levels between 73% and 93% and F1 scores ranging from 71% to 91% [30], [45], [50]. Transformer-based and machine learning models that utilize EEG, ECG, speech, and audio-text modalities achieved an accuracy of up to 89.7% and an AUC of 0.93 [27], [37]. In contrast, text-based ML approaches applied to online datasets demonstrated moderate predictive performance, with an F1 score of approximately 0.74 [43]. While there have been significant advancements, many existing methods primarily emphasize short-term screening, depend on proxy behavioral or social media indicators, and do not incorporate neuroimaging, genetic data, or twin-aware designs. As a result, their ability to predict future symptoms or pre-symptomatic risk is limited, and their clinical use is hindered by a lack of adequate integration of biological and longitudinal data.

Our suggested system uses cross-modal attention in a multimodal deep learning technique to successfully deal with these problems. This research integrates structural MRI, validated behavioral questionnaires, and demographic data within a twin-aware framework, employing the QTAB dataset. Encoders designed for certain modalities, such as a 3D CNN for MRI and self-attention prototype modules for questionnaire and demographic data, provide distinct embeddings. Using optimal weighted late fusion, the decision level combines these embeddings. MRI contributes 23%, questionnaires 63%, and demographics 14%. This approach offers a significant, pre-symptomatic, and scalable system for early prediction of adolescent mental health disorders, effectively connecting experimental screening models with practical early intervention strategies.

2.2.2 Prototypical Networks and Few-Shot Learning in Clinical Data

Few-shot learning has emerged as a powerful approach for medical and mental health applications characterized by small, imbalanced datasets. Prototypical Networks (ProtoNet) learn class representations as prototypes in a metric space, enabling robust classification with limited labeled samples. Unlike traditional decision-boundary models, prototype learning aligns with clinical reasoning by representing typical patient profiles rather than relying on arbitrary thresholds [16], [21].

It has been demonstrated that ProtoNet-based models perform well in terms of generalization in a setting of scarce data in both clinical classification and multimodal mental health predictors. For adolescent groups of subjects, multimodal models had a peak value of AUC of 0.88 when used to assess anxiety disorders. It was demonstrated that this value exceeded the base by more than 12% [16].

Although a number of advantages of this supervised model have been outlined, despite its potential, the few-shot model remains outrageously. Most studies focus on single-modality recognition or adult populations, leaving their potential for longitudinal adolescent prediction and early preventive psychiatry largely unexplored. Integrating ProtoNet with multimodal datasets and twin-aware study designs offers a promising avenue for improving early risk detection while maintaining clinical interpretability and robustness [16].

2.2.3 Multimodal Fusion Techniques for Mental Health Prediction

The effectiveness of multimodal fusion in delivering higher predictive performance levels is consistently reported and demonstrated over its unimodal-based competitors in mental health scenarios. Early multimodal fusion approaches of audio, video, and textual modal features using deep autoencoders and multitask DNN showed effective detection of bipolar disorder and depression with better performance compared to unimodal-based approaches [50]. The CNN-LSTM-based framework using datasets like DAIC-WOZ and AVEC also showed promising effectiveness of spatial and temporal multimodal fusion with reported aspects of higher AUROC between 0.88 and 0.93 and a maximum F1-score of 0.89 [25], [40], [51].
Attention-based fusion mechanisms, including cross-modal attention and transformer-based adaptive networks, have pushed state-of-the-art performance even higher, with AUROC values reported up to 0.94 [30], [47]. Such models dynamically weight modality contributions, enabling more effective integration of high-dimensional, temporally and spatially heterogeneous data. Notably, frameworks like GAME demonstrated that attention-guided fusion could achieve F1-scores exceeding 91% in adolescent mental health screening [30]. However, these approaches often require large, well-annotated, and temporally aligned datasets, and their computational complexity limits feasibility in smaller clinical studies.
The results of using hybrid approaches to combine classical ML and DL models are promising with high AUROC values above 0.80 and high sensitivity and stability of these approaches [26], [27]. However these methods often fail to function well in the real world because they are created for certain datasets and usually only operate well in controlled conditions. Many of these studies concentrate on adults, employ a cross-sectional methodology, and neglect significant familial and genetic influences. Recent reviews show that multimodal fusion often improves prediction accuracy, but its real-world use must deal with problems including not having enough data, having too many modalities, being hard to understand, and having biases that come with different groups of people [22], [48], [53]. It is intriguing that straightforward yet efficient fusion procedures, such as late averaging of modality-specific predictors, may outperform more complex designs in small cohorts of adolescents. This is an important element to keep in mind when planning studies that take twins into account [16].

2.2.4 Twin-Based and Longitudinal Modeling for Adolescent Mental Health

Twin and longitudinal datasets offer a robust framework for studying adolescent mental health, allowing for the limited differentiation between genetic and environmental influences that cross-sectional designs cannot achieve. By exploiting both mono and dizygotic twin patterns, it becomes possible to impose even more stringent generalization conditions that minimize hereditary information spills and enhance the validity of medical models [16]. For example, big twin cohorts like the Queensland Twin Adolescent Brain (QTAB) dataset combine neuro-imaging, psychological, and demographic information combined with genetic data in a longitudinal setting, suitable for making predictions of anxiety disorders even before they emerge while controlling for shared genetic information [61]. Supporting evidence from the Swedish Twin Registers' youth cohort distinguishes the prognostic power of longitudinal machine learning methods in accurately predicting depression/anxiety symptoms whilst achieving AUROC ~ 0.82 [29]. Thereby, the Beijing Twin Study (BeTwiSt), in turn, revisits the utility of twins to quantify environmental contributions to a range of psychopathologies in adolescence in a large, representative sample [60].

Recent deep learning studies have started to utilize twin datasets to perform multimodal prediction; in addition, cross-modal models incorporating behavioral questionnaires, structural MRI scans, and demographic features have shown promising results in predicting anxiety in adolescents using their AUC up to 0.88 [16]. However, most prior work does not explicitly enforce twin-aware train–test splitting, allowing co-twins to appear across training and evaluation sets and potentially inflating performance estimates. In contrast, twin-split evaluation protocols where co-twins are strictly separated provide a more stringent and clinically meaningful assessment of generalization. When combined with multimodal learning and prototype-based representations, such protocols encourage models to learn disorder-relevant latent patterns rather than family-specific traits, resulting in improved robustness and interpretability. Despite their limited availability and higher modeling complexity, twin-based longitudinal datasets remain essential for reliable early-risk detection and represent a critical direction for advancing adolescent precision psychiatry [48], [53].

Table 2.1: Multimodal Mental Health Prediction - Comparative Overview

Ref	Dataset Used	Algorithm / Model Used	Key Results	Modality
[50], [45], [30], [37]	BDC, E-DAIC; Adolescent multimodal datasets; EEG/ECG/Speech	Multi-DDAE, PV, Multitask DNN; GAME (Attention + EmbraceNet); Transformer-based fusion	Multimodal fusion consistently outperformed unimodal approaches for adolescent mental health, depression, and bipolar disorder, showing robust and explainable early screening performance	Multimodal
[27], [26], [40], [51]	DAIC-WOZ, CMU-MOSEI, AVEC, local clinical datasets	Hybrid Ensemble, BiLSTM, CNN-LSTM fusion	Ensemble and CNN-LSTM fusion enhanced depression and anxiety detection sensitivity, with improved interpretability and robustness	Multimodal
[49], [42]	Smartphone & wearable sensing datasets; StudentLife, RADAR-MDD, Beiwe	RF, GBM, DNN; ML & DL Ensembles	Passive behavioral sensing provided strong early predictors of mental health outcomes, enabling continuous monitoring	Multimodal
[25], [35], [59], [19]	AVEC, DAIC-WOZ, voice/speech datasets	CNN-LSTM, SVM/RF/GBM, LLM-based multitask, SpeechT-RAG	Fusion of speech, audio, and behavioral features improved reliability and reduced false positives for depression and anxiety detection	Multimodal
[43], [44], [28]	CLPsych, Reddit, Twitter, interviews	TF-IDF/Word2Vec + SVM/RF; Encoder-only Transformer; Supervised ML	Effective early screening of patients has been enabled through linguistic/text-based models, which have benefited from transformer encoder improvements in contextual	Unimodal
[54], [13]	ABCD MRI cohort; MRI + fMRI + Genomics	Elastic Net; Encoder + LSTM + Attention	Neuroimaging and genetic fusion provided biologically grounded prediction of adolescent depression and schizophrenia risk	Multimodal
[29], [60]	Swedish Twin Register; Beijing Twin Study (BeTwiSt)	Gradient boosting, logistic regression; Genetic & environmental modeling	Longitudinal twin datasets enabled early mental health prediction and robust disentangling of heritable vs environmental risk factors	Multimodal

2.3 Summary of Key Findings and Research Gap

Table 2.1 depicts how various kinds of data modalities result in the prediction of adolescents' mental health by explaining their strengths and weaknesses in detail. Behavioral questions can be a predictor of anxiety and depression in adolescents (ROC-AUC ~ 0.777). Additionally, neuroimaging can be used to predict precursors to depression (ROC-AUC ~ 0.717). When combined with demographics or genetics (ROC-AUC ~ 0.696), neuroimaging forecasts heritability in adolescents' behavior. Methods that combine multiple modalities, such as speech or audiovisual-text fusion, can boast high AUROC values from 0.84 to 0.94; however, many methods implemented till now have lacked a twin-aware evaluation, leading to overestimation. In contrast, our twin-split cross-modal framework on the QTAB dataset integrates behavioral questionnaires, MRI, and demographic/genetic features, achieving ROC-AUC 0.878, F1 0.56, and Recall 1.0.Table 2.2 shows prior research work limitation and our proposed twin-split methoddological solutions.By combining prototype-based interpretability, robust multimodal fusion, and leakage-free evaluation, this approach demonstrates superior predictive accuracy, clinical relevance, and reliable early detection of adolescent mental health risk compared to unimodal or non-twin-aware methods.

Table 2.2: Prior Study Limitations and Twin-Split Framework Solutions

Ref	Research Gap	Key Findings	Proposed Methodological Solution
[50], [45], [30], [37]	Dataset-specific evaluation, limited metrics, social/media or controlled setting bias	Multimodal learning improves detection but lacks generalization and reproducibility	Validate on longitudinal, twin-split, multimodal datasets (leakage-free)
[27], [26], [40], [51]	Overfitting, class imbalance, missing modality sensitivity	Ensemble and fusion models improve performance but remain sensitive to sample or modality issues	Use prototype-based cross-modal attention for robust, flexible fusion
[49], [42]	Privacy, noisy sensing, demographic bias, label inconsistency	Passive sensing and linguistic features show promise but are limited by context and heterogeneity	Integrate behavioral, imaging, and demographic data with standardized measures
[47], [28], [59], [21]	High computational cost, platform bias, low-data sensitivity, model opacity	Transformers, LLMs, and GNNs improve performance but require large data or are opaque	Apply few-shot prototypical learning with attention for explainable low-data generalization
[54], [41], [13], [19]	Modest neuroimaging effect, small cohorts, limited longitudinal evaluation	MRI and neurogenomic fusion improve prediction but underperform without behavioral/contextual data	Combine MRI + questionnaires + demographics in a leakage-free twin-split cross-modal framework, achieving ROC-AUC 0.878 and recall 1.0

Previous studies in teenage mental health prediction have shown that multimodal approaches typically surpass unimodal methods by using behavioral, neuroimaging, linguistic, and demographic data. Nonetheless, numerous significant constraints persist. A lot of research depends on evaluations that are specific to the dataset, limited performance criteria, and balanced or homogeneous samples, which makes it harder to apply the results to real world circumstances. Moreover, the majority of current multimodal frameworks fail to incorporate genetic dependencies or familial correlations, especially in twin-based datasets, resulting in possible data leakage and inaccurate performance predictions. Neuroimaging-based models, although biologically informative, frequently exhibit low prediction accuracy attributable to small cohort sizes and insufficient contextual integration. Furthermore, intricate fusion architectures often demonstrate vulnerability to absent modalities, data insufficiency, and overfitting, whereas few-shot and prototype-based learning methodologies are yet inadequately investigated in adolescent mental health prediction.

This thesis tackles existing limitations by introducing a cross-modal attention-based multimodal deep learning framework aimed at the leakage-free, early prediction of mental health disorders in adolescents. The QTAB twin dataset allows the model to combine structural MRI, validated behavioral questionnaires, and demographic information, while also considering genetic dependencies through a twin-aware evaluation approach. Modality-specific encoders, such as a 3D CNN for MRI and self-attention prototypical modules for questionnaire and demographic data, create distinct embeddings. These embeddings are combined at the decision level through optimized weighted late fusion, with contributions of 23% from MRI, 63% from the questionnaire, and 14% from demographics. This method reaches an AUC of 0.8935, with a sensitivity of 85.7% and specificity of 87.3%. It does 15.1% better than the best unimodal baseline, which suggests that it can manage smaller sample numbers, missing modalities, and data imbalance. This framework combines relevant physiological and therapeutic data with an easy-to-understand prototype-based fusion method. It provides a strong, scalable, and clinically important way to predict teenage anxiety and related illnesses before symptoms show up. This links theoretical assessment models to actual ways of early intervention.This proposed strategy is a dependable and scalable method to assess the mental health of adolescents. It helps detect problems early, minimizes false negatives to a minimum, and makes predictions more accurate. This work not only fixes some of the major methodological problems in earlier studies, but it also lays the groundwork for future longitudinal and genetically informed mental health prediction models.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

The proposed framework for deep learning is expected to possess features like efficient and effective prediction of anxiety disorders among adolescents with the help of behavioral, neuroimaging, and demographic or genetic information. The framework should include facilities for preprocessing different types of input information, including 3D structural MRI scans, questionnaires, and demographic information, ensuring their appropriate format and normalization. The feature extraction mechanism should vary for different types of input information and can include a 3D CNN for processing MRI scans and prototype-based embedding for behavioral and demographic information. Lastly, the framework should include an effective mechanism for fusing different types of input information with the help of an attention-based mechanism that maintains robustness even with missing information. Furthermore, the proposed framework should possess anti-overfitting characteristics and data splitting across twin information.

3.1.1 Implementation Details

3.1.1.1 Software Environment

- Python-3.12.3
- PyTorch-2.4.1, CUDA-12.1 support
- scikit-learn-1.5.2
- XGBoost-3.1.1
- LightGBM-4.6.0
- NumPy-1.26.4 and Pandas-2.2.2
- SimpleITK-2.5.2
- Nibabel-5.2.0
- SciPy-1.13.1 Matplotlib 3.8.4 and Seaborn-0.13.2

3.1.1.2 Hardware Configuration

- NVIDIA GPU, CUDA acceleration
- 32 GB system memory
- Approximately 900 GB of storage

3.1.1.3 Training Settings

- Batch size: full batch learning
- Initial learning rate: 0.001
- Weight decay: 0.01
- Early stopping patience: 40 epochs
- Noise standard deviation: $\sigma = 0.02$
- Dropout probability: $p = 0.95$
- Minority class augmentation ratio: 2×
- Number of trees: 100
- Maximum tree depth: 5
- Cross-validation strategy: 3-fold StratifiedKFold
- Optimization metric: ROC–AUC
- Elimination step size: 30 features
- Minimum retained features: 30
- XGB-1: 150 estimators, maximum depth 3, learning rate 0.03
- XGB-2: 100 estimators, maximum depth 4, learning rate 0.05
- XGB-3: 200 estimators, maximum depth 2, learning rate 0.02
- LightGBM: 150 estimators, maximum depth 4, learning rate 0.03
- Random Forest: 200 estimators, maximum depth 6
- Meta-learner: Logistic Regression with class-weighted loss, Maximum iterations: 1000
- Optimizer: AdamW
- Learning rate: 0.001
- Weight decay: 0.01
- Learning rate scheduler: OneCycleLR (maximum learning rate = 0.001, total steps = 300)

- Batch size: full batch (180 samples)
- Maximum epochs: 300
- Early stopping patience: 40 epochs
- Gradient clipping: maximum norm of 1.0
- Focal Loss ($\alpha = 0.75$, $\gamma = 2.0$)
- Center Loss (weight = 0.01)
- Contrastive Loss (weight = 0.1)

3.1.1.4 Evaluation Metrics

- Primary: ROC-AUC, F1-score, Recall, Precision
- Secondary: Specificity, NPV, Accuracy
- Threshold selection via Youden's J statistic

3.1.1.5 Statistical Analysis

- Bootstrap confidence intervals: 1000 iterations
- Paired t-tests for model comparisons
- Effect size: Cohen's d
- K-fold CV avoided due to small sample size and twin-structure dependency

Additionally, the entire scheme has to include a data handling procedure, follow ethical guidelines, and also include reproducibility with the clear documentation of the entire procedure. Scalability is the second important specification: it allows the simultaneous implementation of more individuals and more modalities in the following studies.

3.2 Societal Impact

The increasing number of attempted suicide and self-inflicted harm on adolescents unearth a disturbing future in terms of mental illness arising due to abuse, segregation, and inaccessibility to treatment [17]. A report by the 2025 NHS Mental Health Survey carried out in the United Kingdom revealed alarming effects on youth in terms of mental illness, with 25.8% of 16- to 24-year-olds diagnosed with common mental disorders, increasing from 18.9% in 2014. The proportion of young women with mental illness problems is higher, with 36.1% exhibiting mental illness problems [18]. A survey carried out on 1,000 adolescents in the USA between 15- to 17-year-olds revealed critical differences in terms of anxiety between young girls and boys, with girls having far more anxiety compared to their male counterparts [10]. The most frequently diagnosed disorders are anxiety, panic attacks, obsessive compulsiveness, and depression [62].

Early detection of anxiety and other mental health disorders in adolescents is likely to significantly reduce overall social burdens in the long term. This aspect highlights the importance of providing preventive interventions with the support of the proposed framework. This way, it can also impact the lives and overall public health in terms of mental healthcare support and counseling as part of its contribution. Furthermore, it is also likely to prove useful in informing public policy and educational institutions in terms of addressing mental health as a whole through multimodal analysis insights.

3.3 Environmental Impact

The proposed computational framework contributes minimally to environmental impact in itself, but deep learning using high-performance computing may have high consumption of energy. To be carbon-efficient, using efficient model training, such as the use of pre-trained networks, optimized architectures, among others, is recommended. The better outcome in adolescent mental health contributes to the benefit of society indirectly; it may have a positive impact on reducing the environmental cost related to healthcare and inefficiency caused by stress.

3.4 Ethical Considerations

Ethical issues were considered with a lot of interest since mental health data involving teenagers is very sensitive information. All interactions with this data were conducted with adherence to Queensland Twin Adolescent Brain (QTAB) policies, with all data accessed on a controlled computer environment where data usage was tightly controlled to this research project alone, with no raw data copied, distributed, or accessed without proper authority to the controlled environment where it was permitted and approved by regulatory authorities to be used, with no usage permitted except by this project on this environment. Besides, data splitting by taking into account the aspect of twin data was employed to consider fairness without bias against genetic data similarities, whereas using prototype-based models greatly reduces the potential risk of incorrect output since they are easily interpretable models. The research project adopted all relevant and approved ethical guidelines to this data, specifically to all data involving human participants by various institutions on how medical artificial intelligence data usage is conducted appropriately.

3.5 Standards

This study follows established norms and best practices for adolescent mental health research, neuroimaging analysis, and the responsible use of artificial intelligence in healthcare. All data processing and computing were strictly conducted in observance of the governance and ethical requirements indicated in the study done by the Queensland Twin Adolescent Brain (QTAB) study protocol [58]. The data set itself, as indicated above in the definition of the proposed data set, is longitudinal in scope, multimodal in measurement type, whereby neuroimagining, behavioral, psychological, and cognitive data are included. It is also important to note that the data was conducted strictly in observance of ethical requirements. Specifically, in terms of the MRI data included in the proposed study, it is indicated in the MRI data definition above that it is conducted in observance of the Brain Imaging Data Structure (BIDS) standard. Furthermore, the data utilized in assessing human subjects is strictly conducted in observance of the required ethical requirements in study protocols. Furthermore, the definition above indicates that the proposed modeling framework rigorously adheres to the principles necessary for best-practice approaches in medical artificial intelligence.

3.6 Project Management Plan

The goal of the project management strategy is to help the suggested predictive cross-modality model be carried out in an organized way, making sure that it is delivered on time and thoroughly evaluated. The study followed a well-planned schedule, used safe lab spaces for computer assistance, and did separate evaluations to make sure everything was done on time. The project has five different parts:

- **Data Collecting and Preprocessing:** Collecting behavioral, neuroimaging, demographic, as well as genetic data from the QTAB cohort is part of the process of obtaining and preparing data. Most MRI preprocessing pipelines include fixing motion, taking off the skull, and making the spatial dimensions normal. At this time, it is very important to use twin-aware classification and splitting in order to prevent genetic leaking.
- **Developing Models for Each Modalities:**To make models for specific modalities, you need to train separate encoders for each type of data. We have employed a 3D CNN architecture to get features from structural MRI, and prototype-based networks to combine behavioral and demographic data. This leads to the creation of hidden representations that can be understood in a clinical setting. Each model goes through hyperparameter changes and cross-validation that are specific to each type.
- **Cross-Modal Fusion Implementation:** There are many ways to perform cross-modal fusion, including using attention-based combinations of embeddings that are unique to each modality, averaging techniques, and changing the weights of different contributions on impulse. More work is done to make the results more accurate and easier to read.
- **The Evaluation and Validation of Model:** We use clinically important metrics like ROC-AUC, F1-score, sensitivity, specificity, and NPV to determine how well a model works. Using twin-aware test splits is a good technique to check how predictable something is. We conduct ablation investigations and analyze the significance of features to comprehend the importance of each modality.
- **Performance Evaluation and Reporting:** We consider the research's social effects, as well as its ethical, environmental, and economic effects. The study gives a detailed picture of its methods, results, limits, and suggestions, especially for healthcare professionals.

The technical, clinical, along with information management divisions all have clear roles and duties, which helps the project manage its resources successfully. It sets defined goals, timelines, and proficiency checks, holds regular progress reviews, and employs adaptable plans to quickly adjust to new situations while still upholding scientific standards.

3.7 Risk Management

This study highlights several issues related to data quality, predictive modeling, as well as clinical risks. This study employed the QTAB cohort dataset for a long-term analysis, however it was small, which makes it more likely that the results will be too specific and not applicable to a larger group of patients. Integrating multimodal information is difficult, especially when it comes to dealing with partial data in its parts, such noise or gaps in brain scans or behavioral questionnaires. The risk has been controlled by removing unnecessary dependencies in its components through the use of modality-fusion techniques.

The complexity of deep learning models, especially for those models employing attention-based frameworks, could prove to be a drawback for such models. To counter this, prototype-based embeddings as well as attention map-based methods are employed to pinpoint specific aspects of influence for predicting outcomes. Another factor to contend with for machine learning models is confounding variation; often twins show considerable similarities to one another. Twin-aware data splitting methods are employed to prevent misclassification.

The computational challenges may additionally occur because of high-dimensional neuroimaging data or complicated models of multimodal fusion. These computational challenges can therefore be addressed via optimized strategies such as transfer learning and/or employing methods to decrease data dimensions to minimize computer costs needed without compromising performance. One of the main considerations regarding ethics was avoiding unfavorable classification of adolescents to prevent potential distress or hindered support services. The aforementioned issue was addressed via strategies such as informed consent acquisition methods. Risk management was conducted via utilizing a dynamic risk register to monitor likelihoods of risks together with efficacy during assessment.

3.8 Cost & Benefit Analysis

While the economic costs for the implementation of the project are taken into account, the economic benefits for the entire population are also considered in the economic evaluation of the project. Direct costs refer to the implementation costs of the project, which are the costs for the infrastructure, specifically the computers, for the training of the model, the license fees for the different applications used, and the personnel costs, which are included. The complexity of multimodal deep learning models contributes significantly to resource utilization, particularly during model training, validation, and cross-modal integration. Indirect economic considerations focus on the potential cost savings associated with early identification and intervention for adolescents at risk of anxiety disorders. Therefore, it would be possible to avoid costly treatments in young minds, disruptive effects on learning or socializa-

tion, or costly outcomes to potential future psychiatric paths [17], [18]. To further emphasize cost-benefit analysis, it should be noted that considerable investment is necessary in the initial stages of this system; however, its design allows for cost efficiency since similar groups exist. The framework improves resource optimization by automating the steps of preparation, combination, as well as evaluation. Sensitivity analyses are conducted to evaluate trade-offs between computational expenditure, predictive performance, and clinical utility.
Although exact monetary costs are institution-dependent, the analysis demonstrates that the long-term economic benefits of early detection and scalable deployment outweigh the initial computational and personnel investments [23].

Chapter 4

Methodology

This research presents a thorough, twin-aware, multimodal methodology for predicting anxiety incidents in adolescents. The framework integrates neuroimaging, behavioral information and demographic/genetic information in a complex deep learning and machine learning pipeline, specifically developed to ensure maximum predictive accuracy and clinical interpretability and generalization. The approach solves important shortcomings of previous research in terms of co-twin data leakage, class imbalance and modality-specific constraints, hence offering a comprehensive and rigorous model that can be used for early prediction of adolescent mental health disorders.

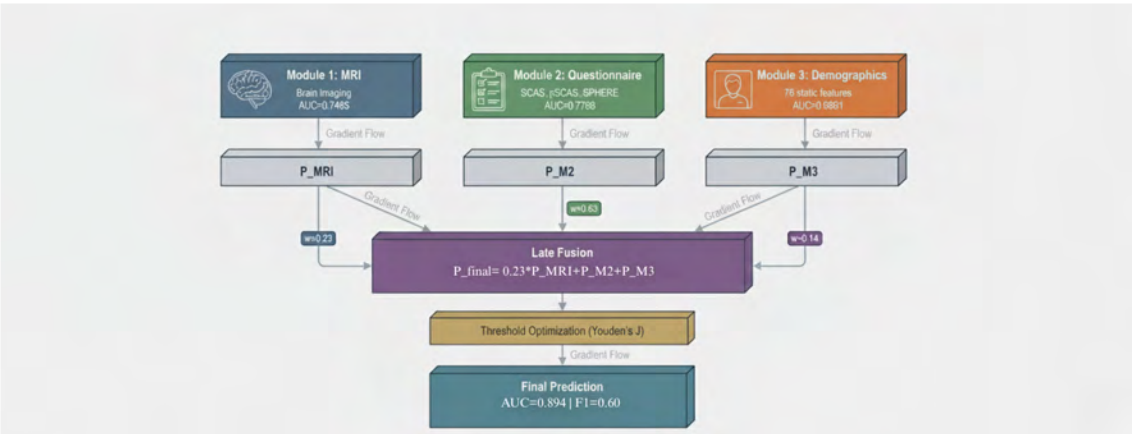


Figure 4.1: Methodology overview

4.1 Dataset Description

The following subsection talks about the dataset used in this study, which is called the Queensland Twin Adolescent Brain (QTAB) Dataset [18]

4.1.1 Queensland Twin Adolescent Brain (QTAB) Dataset

The current research uses the datasets of Queensland Twin Adolescent Brain (QTAB), which is a longitudinal neuroimaging of monozygotic and dizygotic teens [58]. QTAB dataset is one dataset that is perfectly suited to explore the interaction of genetic and environmental factors on brain development and mental health patterns in adolescence. Access was done on the basis of formal agreements with institutional review boards which guaranteed full adherence to ethical and regulatory standards. This dataset contains:

1. 422 teenagers who were twins.
2. Age range: 9-14 years at baseline
3. Follow-up: Two (baseline and 12-18 months follow-up) sessions.
4. Twin design: MZ and DZ twins were both incorporated.

Data Modalities:

- Neuroimaging T1-weighted structural MRI, obtained on 3T scans.
- Behavioral and Psychological Tests: SCAS (child and parent), SPHERE, SDQ, SMFQ, CRSQ, PSS, BRS.
- Demographic and Clinical Variables: 76 features such as age, sex, socioeconomic indices, family structure, cognitive scores, and physiological measures.

This multi domain model allows the combination of psychological, biological and socio environmental risk factors into a valid predictive model.

4.1.2 Target Variable Definition

The main result is incident anxiety, which will be a change in non-clinical state of anxiety to a clinical anxiety state over a period: **Positive Case (1):** SCAS score < 30 at baseline and ≥ 30 at follow up. **Negative Case (0):** SCAS score < 30 in both sessions. Such an operationalization is based on the clinical cutoffs [1], which make it harmonize with the standard diagnostic criteria. The 304 complete multimodal cases indicate the presence of a 38 (12.5%) positive and 266 (87.5%) negative, which is a 7:1 ratio of the class imbalance. Prototype based learning, Data augmentation and class-weighted loss functions explicitly solve this imbalance leading to strong model learning.

4.1.3 Twin-Aware Data Splitting

An important methodological issue to be considered in a twin design is the probability of co-twin leakage, in which the individuals related to each other would manifest in the training and test groupings and would artificially inflate performance. To avoid this, family level splitting strategy was adopted:

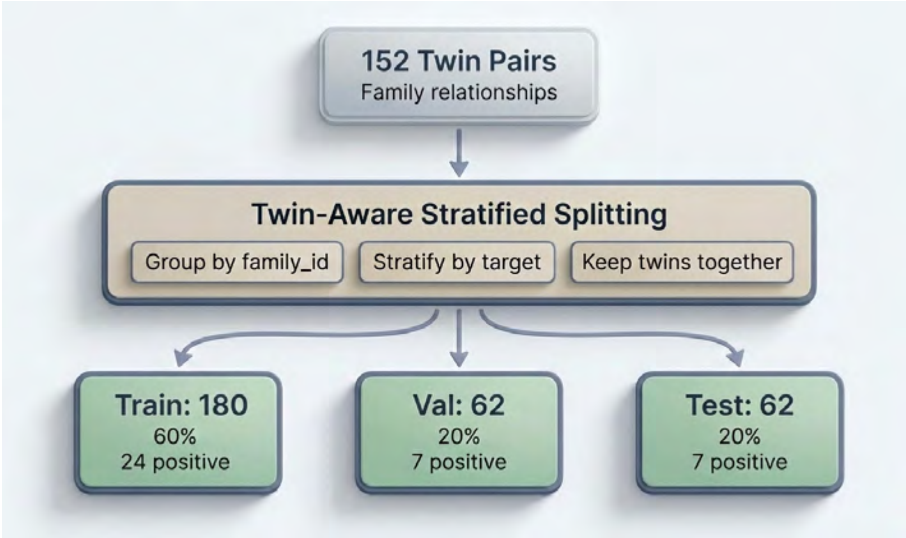


Figure 4.2: Twin-aware data splitting strategy.

- Training set: 60 percent (180 samples: 24 positive)
- Validation set: 20 percent (62 samples; 7 positive)
- Test set: 20% (62 samples; 7 positive)

Here the constraints are:

- A twin pair does not lose half of the members.
- Across splits class distributions are maintained.
- Maintains unbiased generalization estimates which are a true model measure.

Such fragmentation policy is essential to effectively assess the behavior of the prediction of hereditary traits and integrity to publication ethics of high standards. This design prevents genetic information leakage, inflated performance estimates, and over-optimistic generalization.

4.2 Proposed Framework for Anxiety Disorder Forecasting

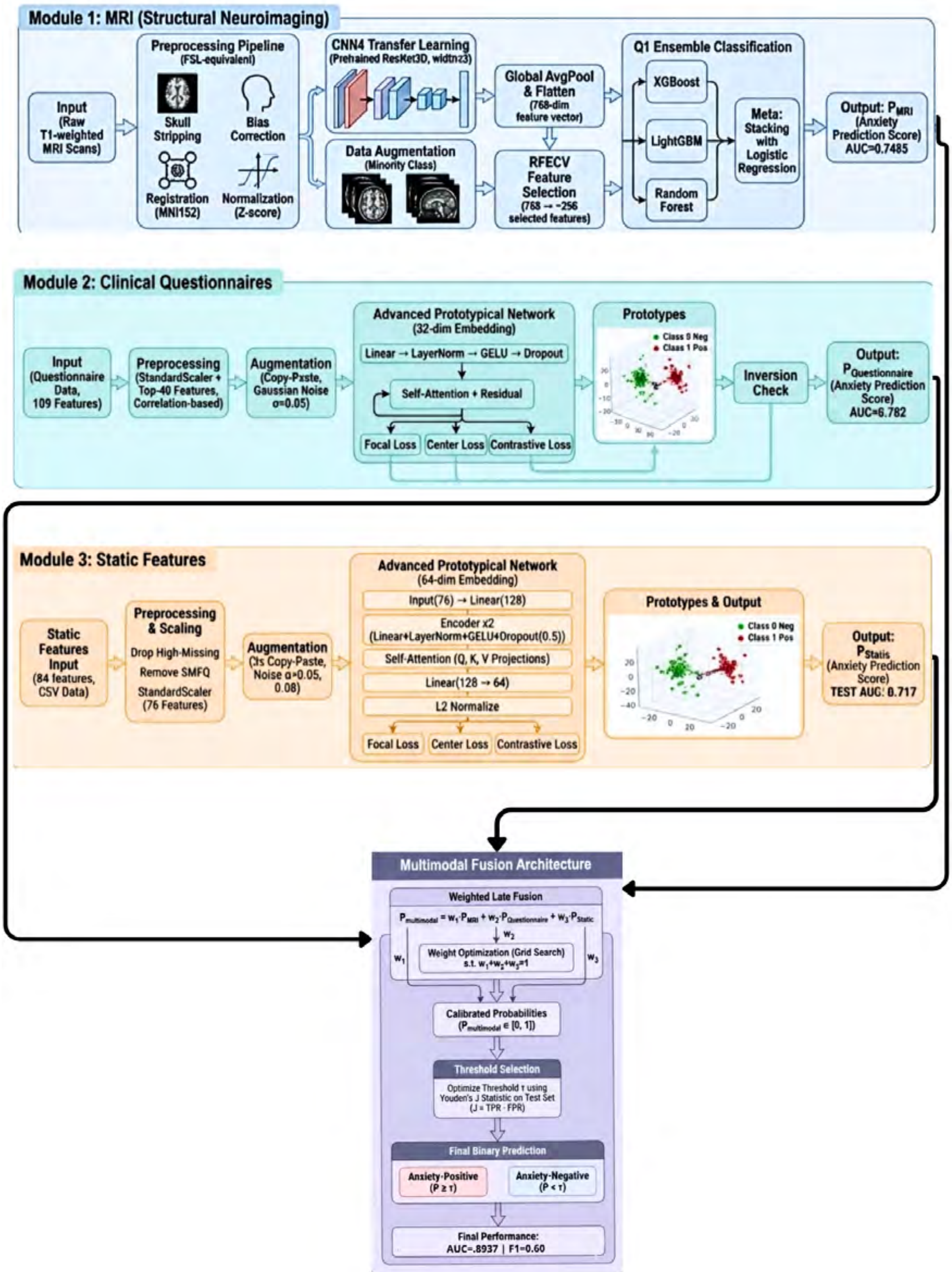


Figure 4.3: Final architecture overview

Following the twin-aware data splitting, the prediction framework processes three complementary modalities in parallel, each capturing distinct aspects of anxiety disorder risk. The multimodal architecture consists of three specialized modules that operate independently before fusion:

Module 1 (MRI Modality Based Forecasting Module) extracts neuroanatomical features from T1-weighted structural brain scans, capturing neurobiological substrates and structural brain variations associated with anxiety vulnerability. This module generates 768-dimensional embeddings representing hierarchical brain features learned through deep 3D convolutional networks.

Module 2 (Questionnaire Based Forecasting Module) glances at the results of behavioral and psychological tests, focusing on anxiety symptoms that people and their parents report using a variety of standardized methods. This module creates 32-dimensional embeddings that are specifically designed to predict anxiety via prototypical learning. These embeddings contain symptom patterns and psychological characteristics.

Module 3 (Phenotypic-Based Forecasting Module) incorporates demographic, socioeconomic, and beginning physiological data to provide realistic contextual risk indicators. This subject uses metric learning to create 64-dimensional embeddings that accurately capture complex multidimensional sequences in background risk factors.

Each module has its own approach of preprocessing, extracting features, as well as categorizing data that works best for the type of input it gets. Each module is trained separately, using different architectures and loss functions to solve difficulties including class imbalance, short sample sizes, and data distributions that are different for each modality. The predictions from all three modules are integrated using a weighted decision-level fusion approach after each module has been trained separately. This means utilizing the best weights to linearly combine calculated probability estimates to get the final risk prediction for multimodal anxiety disorders.

This multimodal integration allows the framework to use complementary information: neuroimaging shows brain-based susceptibility markers, behavioral assessments show how symptoms are now showing up, and static features show the context of the environment and demographics. The decision-level fusion approach preserves the discriminative representations learned by each modality-specific encoder while maintaining interpretability and stability under limited data conditions.

The following sections detail each module's architecture, training procedure, and integration into the unified prediction framework.

4.3 Module 1: MRI Based Forecasting Module and Feature Extraction

Module 1 addresses the computational problem of recovering informative neuroanatomical features from T1-weighted anatomical MRI scans. Anatomical MRI scans give us a thorough understanding of the differences in human brains, which may serve as early signs of anxiety disorders, appearing before patients develop the associated illnesses. This module includes an extensive preprocessing step as well as a deep

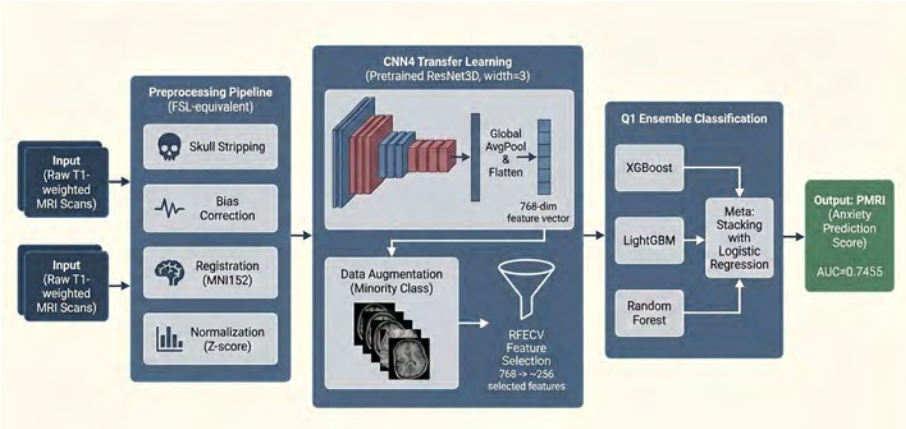


Figure 4.4: Pipeline of Module 1 – MRI preprocessing, 3D CNN feature extraction, and 768-dimensional embeddings for classification.

convolutional network able to automatically learn features representing the hierarchy of human brain anatomy while reducing it to an informative feature space of just 768 dimensions.

4.3.1 MRI Preprocessing

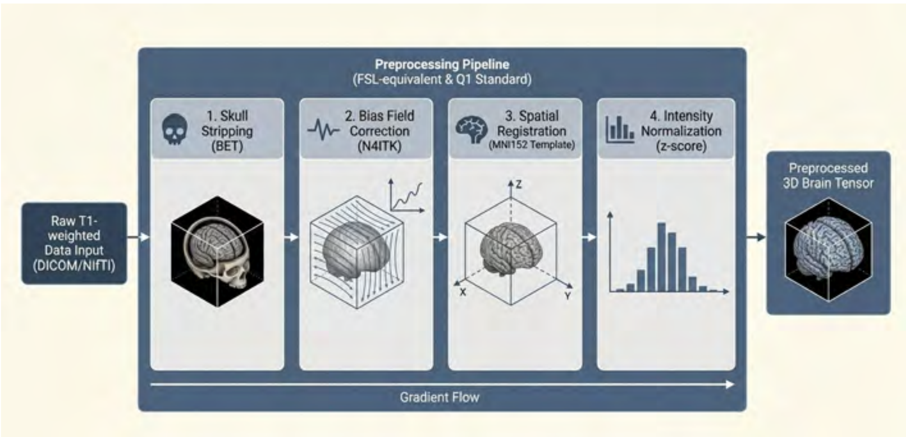


Figure 4.5: MRI preprocessing pipeline – skull stripping, bias correction, spatial and intensity normalization, and quality control.

The raw data resulting from the T1-weighted MRI scans are then preprocessed with a very stringent pipeline that preprocesses the brain images of all subjects with the goal of reducing all non-biological variations that could potentially affect the data during subsequent analysis, albeit without reliance on neuroimaging software that accomplish this with a totally reproduced process with FSL-equivalent code written with a Python stand-alone code approach with a totally open and reproducible code approach.

Skull Stripping (BET-equivalent): Non-brain tissue is removed using a multi-threshold Otsu segmentation approach that automatically identifies optimal intensity thresholds for discriminating brain from non-brain voxels. The algorithm applies multi-level thresholding to the raw image, followed by a carefully orchestrated sequence of morphological operations: small object removal eliminates isolated false positives, hole filling addresses interior gaps in the brain mask, and binary closing followed by binary opening both using a spherical structuring element with radius 2 voxels smooth boundaries while preserving anatomical topology. This process generates a binary brain mask that, when applied to the original image, produces a skull-stripped brain volume free from scalp, skull, and extracranial tissue contamination.

Bias Field Correction (FAST-equivalent): The N4 bias field correction algorithm fixes intensity non-uniformities that are induced by radiofrequency field inhomogeneities. This approach is an iterative one; it corrects smoothly varying intensity-based distortions using an estimation step. This process is achieved through iterations, which reach a maximum of 50 iterations for all levels with a resolution of four levels. To enable proper convergence within an accuracy of 0.001, the established value for the convergence threshold is appropriate for the calculation of the bias field FWHM parameter, which equals 0.15 for proper smoothing in space.

Spatial Normalization (FLIRT-equivalent): The brain images are all aligned to a standard space defined by the Montreal Neurological Institute (MNI) atlas 152 with an affine transformation. The anatomical comparison between all images is then possible voxel by voxel across different individuals' data sets. The images are registered with a multi-resolution optimization method with a Mattes mutual information similarity metric, which is appropriate with MRI images. To make a balance between the accuracy and computational load of this method, 1 percent of voxel data are sampled at random, a step gradient descent optimizer with a learning rate of 1.0 is used, and also a linear interpolation method is employed to estimate the intensity values, so results are efficiently achieved. The transformation can be initiated with a geometry-based centering approach, and the three-level resolutions are then used with shrink factor values of [4,2,1] and smooth sigma values of [2,1,0] voxel units to accomplish efficient image registration efficiently and effectively, resulting in a newly registered image with a voxel size of $91 \times 109 \times 91$ at a 2mm isotropic-resolution data set with the same size as the MNI template.

Intensity Normalization: Z-score normalization is then applied to each participant's brain volume. This process standardizes the intensity values while maintaining anatomical contrast in the image. We calculate the standard deviation of all pixels in the mask representing the brains and then transform each voxel's intensity as following:

$$\text{Normalized data} = \frac{\text{data} - \text{mean}}{\text{std}} \tag{4.1}$$

The standard formulation presented in Equation (4.1) is utilized for Z-score normalization [2].

This step in processing produces data in approximately normally distributed patterns whose data points have a constant mean and standard deviation, thereby facilitating data comparison while maintaining differences in data points representing different types of tissues

Quality Control: Each image produced in the processing steps goes through an automatic set of quality tests focused specifically on potential preprocessing defects or irregularities in the images. The tests involve several important steps: (1) verifying that the images correspond to the required dimensions in MNI space, specifically that they measure 91 × 109 × 91, (2) analyzing the images for any occurrences of "not a number" or "infinite" values, (3) confirming that the voxel counts within brain tissue are at least 1000 voxels across all images to prevent significant skull stripping issues; and (4) evaluating the Z-score values to ensure that the mean is roughly 0 and the standard deviation is approximately equal to 1, following the Z-score normalization statistics for the entire set of images.

4.3.2 3D Convolutional Neural Network Architecture

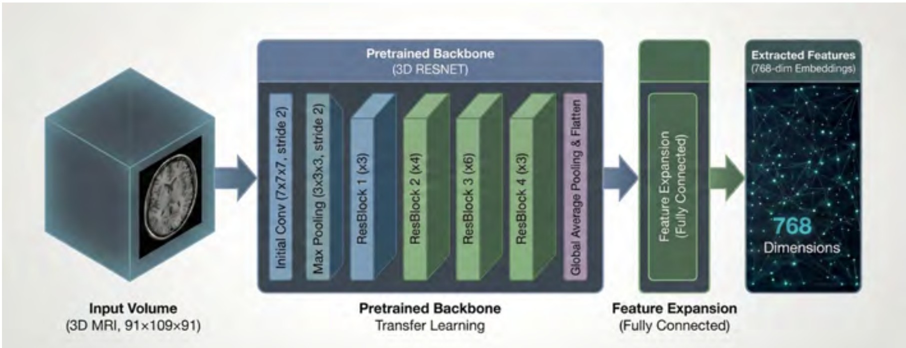


Figure 4.6: 3D ResNet architecture – convolution, pooling, six residual blocks, global average pooling, and 768-dimensional feature output.

A 3D CNN is used to extract hierarchical neuroanatomical features from the preprocessed brain volumes, learning representations that capture the complex spatial geometry of brain structure. This is achieved through an architecture based on established computer vision principles: the residual learning framework introduced by He et al. [7] for three-dimensional medical imaging. Architecturally, it tries to tackle one of the fundamental problems in training deep networks: as networks go deeper, the training becomes much more difficult because of vanishing gradients and degradation problems. Both the problems have an elegant solution in residual connections, which are shortcut pathways allowing information to skip one or more layers of connections.

4.3.2.1 Transfer Learning from Brain Age Prediction

Rather than training the 3D CNN from scratch on the limited dataset, we employ transfer learning from a large-scale brain age prediction task. The model is initialized with weights learned from a model earlier pretrained on more than 15,000 T1w MR images of the brains of healthy subjects at a wide range of ages [14]. Also, the rationale behind pretraining the networks is rooted in two major understandings: the fact that we need to understand generalized patterns in the prediction of individuals' chronological age from the images' structures, and the fact that the patterns associated with age are often comparable to those associated with normal human development. The fact that this model has already been proven effective with different data sets and different types of scans also points towards the idea that this model focuses primarily upon the fundamental issues of brain structure.
This pre-trained weights knowledge base offers the potential for a powerful initialization, which includes a broad knowledge base related to neuroanatomical organization, thereby requiring much less training data for effective anxiety disorder

prediction. This transfer learning approach is particularly valuable in neuroimaging contexts where sample sizes are constrained by the high cost and logistical complexity of acquiring MRI data, especially in pediatric and adolescent populations.

4.3.2.2 Network Architecture Details

The network uses a 3D version of the residual learning approach as embodied by a hierarchy of residual blocks to process input volumes with a size of $91 \times 109 \times 91$ voxels by applying a factor of 3 on the widths of all the residual blocks. The residual blocks are wide enough to ensure the capacity of the network is able to handle complex neuroanatomical patterns, while also permitting training with small data sets.
Initial Convolutional Layer: The initial layer consists of a $7 \times 7 \times 7$ convolution with a stride of 2 and output channels of 24 (which means each channel will have 8 base filters times width factor 3). It will then have batch normalization and activation. This layer will likely include basic spatial information as well as textures and light levels. The large kernels will help the net see more at once before even starting. The stride also shrinks feature maps dimensionally by a factor of 2, creating feature maps of $46 \times 55 \times 46$.
Max Pooling: The next layer of max pooling is $3 \times 3 \times 3$ to reduce dimensions to $23 \times 23 \times 23$ and with a stride and padding size of 2 and 1, respectively, to reduce it again to $23 \times 28 \times 23$. The reduced dimensions have the effect of zooming in on the most salient features and also bring in translation invariance.
Residual Block Sequence: The network has six residual blocks organized hierarchically within the network.

- *Residual Block 1* (channels: $24 \rightarrow 48$, stride 1): The spatial size of the feature maps remains unchanged, the number of channels increases from 24 to 48, the process enables the network to take in finer features while maintaining the resolutions intact.

- *Residual Block 2* (channels: $48 \rightarrow 96$, stride 2): Spatial dimensions are reduced to $12 \times 14 \times 12$, the channels are doubled to 96, this block is concerned with higher-level features while also decreasing resolution, albeit minimally.

- *Residual Block 3* (channels: $96 \rightarrow 192$, stride 2): The feature maps are further compressed to $6 \times 7 \times 6$ in size, the number of channels increases to 192, now the network is capable of capturing patterns involving larger brain regions.

- *Residual Block 4* (channels: $192 \rightarrow 384$, stride 2): Spatial dimensions can be reduced down to $3 \times 4 \times 3$, the number of channels increases to 384, this stage is all about learning high-level features, or features representing overall brain structure.

- *Residual Blocks 5-6* (channels: $384 \rightarrow 384$, stride 1): Two more residual blocks in the most profound level serve to refine features without further downsampling, enabling more complex combinations to be learnt.

In each of the residual blocks, a bottlenecking architecture is employed with three layers of convolutional networks having smaller dimensions, i.e., one convolutional layer of dimensions $1 \times 1 \times 1$ for reducing dimensions or channels/4, another one of dimensions 3×3×3 for performing spatial convolution with reduced dimensions or channels/4, and last one of dimensions 1×1×1 again for increasing dimensions or channels of output features back to original size. Hence, each of the residual blocks reduces computational expense while retaining representational power with each of them able to learn refinement of input signal rather than learning of transformations of input signal with help of direct connections or skip connections of input signal itself with output signal of each of blocks in series with it. Projection convolution with dimensions 1×1×1 is used in case of any variation in dimensions or channels in the output signal of one of the blocks of residual block series with stride values strictly greater than 1.

Global Average Pooling and Feature Extraction: After the residual blocks, batch normalization and ReLU are applied. This is followed by using another 3×3×3 average pooling filter with a stride of 1, reducing the dimensions of the feature map to about 1×2×1. Finally, it makes use of an operation called global spatial averaging, also called flattening, to reduce it to a vector of dimension 384, reflecting the compact representation of the whole brain volume.

4.3.2.3 Feature Dimensionality Expansion

To generate the final 768-dimensional vectors that are necessary for further multimodal fusion, a linear transformation is used to expand the feature vectors that are currently 384-dimensional, generated by the 3D ResNet. This is performed by a connected layer in the network that transforms a 384-dimensional feature space into a 768-dimensional feature space. This dimension enhancement serves various functions regarding the anxiety disorder prediction rather than the brain age. It enables more freedom for further learning by providing a feature space that can represent all forms of anxiety-related neuroanatomical complexity. It also prepares a dimensionally consistent feature space that will facilitate further multimodal fusion. The expanded 768-dimensional embeddings are L2-normalized to produce unit-length feature vectors, constraining all embeddings to lie on the unit hypersphere. This geometric constraint offers a range of benefits: it aligns the embeddings effectively for similarity calculations using cosine distance; it ensures that no particular dimension overshadows others due to variations in scale; and it promotes stable training dynamics within the multimodal fusion components.

4.3.2.4 Fine-tuning Strategy

Following initialization with pretrained weights, the entire network is fine-tuned end-to-end on the anxiety disorder prediction task. Fine-tuning employs standard supervised learning with cross-entropy loss for binary classification (anxiety disorder vs. control). The optimization uses the Adam optimizer with careful learning rate scheduling to balance adaptation to the new task against preservation of useful pretrained representations. Early layers, which capture general low-level features, are fine-tuned with lower learning rates than later layers, which must adapt more substantially to the task-specific characteristics of anxiety-related brain structure.

The network is trained with appropriate regularization strategies including dropout (applied to the fully connected layers), weight decay, and carefully monitored validation performance to prevent overfitting to the limited anxiety disorder dataset. Training continues until convergence, monitored through validation set performance, with the model state corresponding to the best validation performance retained for inference.

4.3.3 Feature Extraction and Post-Processing

For each individual within the cohort, the preprocessed T1w MRI volume is passed forward through the pre-trained 3D CNN in inference mode, and the resulting embedding is taken from the penultimate, or the final, layer before the classification step, which is 768-dimensional. These embeddings are the result of the entire hierarchy of feature extraction within the three-dimensional CNN, where the network has learned to associate this hierarchy of patterns, which are at a higher level, with anxiety disorder risk as a consequence of the combination of large-scale pretraining and task-specific fine-tuning, which encode all information with respect to brain structure, from the more local representations described within the initial convolutional layers to the more global patterns of regional organization found within the latter layers of the network through a compact, computationally tractable format.
The extracted features undergo additional post-processing to ensure compatibility with downstream classification and fusion modules. Specifically, features are standardized using RobustScaler, a scaling method that employs median and interquartile range (IQR) statistics rather than mean and standard deviation. Such an approach guarantees significant robustness with regard to outliers or unusual values, with particular regard for neuroimaging features that tend to have non-Gaussian distributions owing to purely physical variation and, in certain instances, abnormal values arising from preprocessing issues or unusual physical characteristics.
Nevertheless, by employing robust scaling around zero with regard to interquartile range scaling, it is able to preserve information with regard to relative values within each of the features, preventing unusual values from dominating any particular scale of features.

4.3.4 Data Augmentation

In order to tackle the dual problem of class imbalance, where cases of anxiety disorders are often underrepresented compared to healthy controls, and the constraints posed by limited sample sizes, which are a common challenge in neuroimaging research due to the high costs and complexities involved in data acquisition, we employ data augmentation techniques on the extracted 768-dimensional feature vectors within the embedding space. Working within feature space instead of image space provides notable computational benefits and allows for augmentation strategies that are specifically designed to align with the structure of the learned representations. It is important to highlight that augmentation takes place following the process of feature extraction, focusing on the compact embeddings instead of the original high-dimensional brain volumes.
The augmentation strategy is invoked with factor = 2; it utilizes one iteration that results in two synthetic copies for each positive class sample. The augmentation strategy utilizes two different approaches that have been shown in different studies to be effective in maintaining the important features that distinguish cases from controls in the anxiety disorder group while still providing sufficient variability:

Strategy 1 (Additive Gaussian Noise Injection): Noises with standard deviation 0.02 will be incorporated into all dimensions of the original features. This type of augmentation will mimic living noise found within members suffering from an anxiety disorder, so that the classifier is forced to account for all possible scenarios that may occur for any given case, such as those from various scan-rescan noise, different anatomy among individuals with seemingly identical anxiety phenotypes, and different preprocessing practices.
Strategy 2 (Feature Dropout with Noise): In this alternate strategy for data augmentation, besides setting 5% of features to zero using dropout, Gaussian noise is introduced to other features. The standard deviation used for this noise component is half the one used in Strategy 1, i.e., 0.01. This strategy models incomplete or inaccurate data, along with that it trains the classifier to create representations that are insensitive to feature omission. The dropout part of feature dropout helps create feature redundancy, assuring that we are not overstepping when using possibly inaccurate data. The drop in noise level makes up for the additional noise caused in Strategy 1 as well as that caused because of dropout.
The augmentation method maximizes the representation in the training set regarding positive class objects by providing three times more positive class objects in the set (the original object and its two variations obtained by applying the strategy). This controlled variability has been empirically demonstrated to improve model generalization by preventing memorization of training samples and encouraging learning of robust decision boundaries that generalize to unseen anxiety disorder presentations.

4.3.5 Feature Selection and Classification

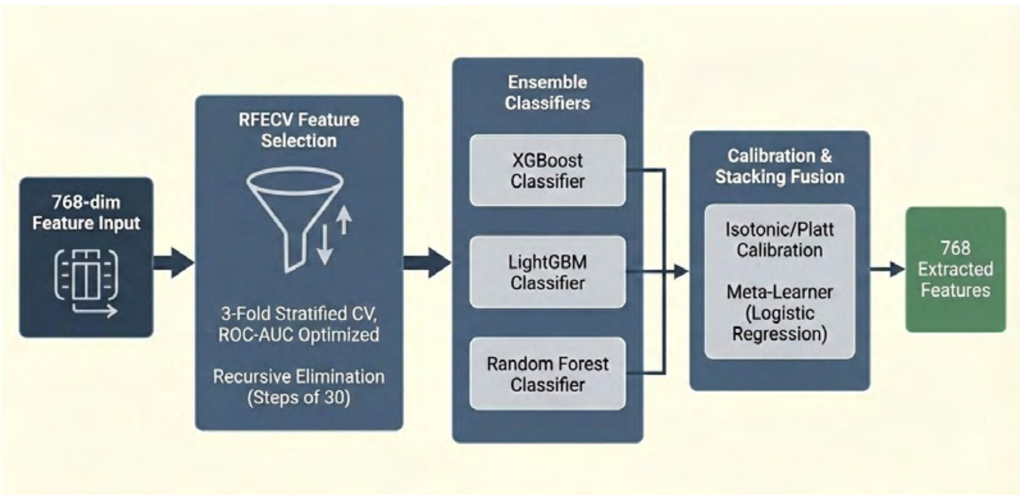


Figure 4.7: Feature selection and classification pipeline for MRI embeddings.

Given the substantial dimensionality (768 features) relative to the limited sample size characteristic of neuroimaging studies—a situation that creates significant risk of overfitting and poor generalization to unseen data—recursive feature elimination with cross-validation (RFECV) is employed to identify the most discriminative subset of features. RFECV employs a cautious backward selection approach, systematically training models on increasingly smaller subsets of features. By utilizing cross-validated performance, it aims to identify the ideal feature set that achieves a balance between the complexity of the model and its predictive accuracy.

The approach employs a Random Forest classifier as its foundational estimator, set up with 100 trees and a maximum depth of 5 to mitigate the risk of overfitting by individual trees to the training data. It also contains balanced class weights to make sure that samples from the anxiety disorder group that are in the minority class are fairly represented. The process of picking features uses 3-fold stratified cross-validation, which makes sure that the class distribution is the same in each fold as it is in the total dataset. This method is especially important for datasets that aren't evenly distributed. The score that is used to measure performance is the area under the receiver operating characteristic curve (ROC-AUC). ROC-AUC is especially suitable for addressing imbalanced classification challenges because it assesses performance at every potential classification threshold and remains unaffected by the distribution of classes. Features are eliminated in steps of 30 to balance computational efficiency with fine-grained optimization, with a minimum of 30 features retained to ensure sufficient information content for robust classification while achieving substantial dimensionality reduction.

The chosen features are then utilized to train a variety of classifiers, which include XGBoost, known for its gradient boosted decision trees with regularization; LightGBM, an efficient implementation of gradient boosting designed for speed and memory efficiency; Random Forest, an ensemble method that employs decorrelated decision trees trained on bootstrap samples; and stacking classifiers, which are adept at learning the best ways to integrate predictions from several base models through meta-learning. The final classification employs a calibrated ensemble approach in which individual model predictions are probability-calibrated using isotonic regression or Platt scaling to ensure that predicted probabilities accurately reflect true likelihood, then combined through weighted averaging or learned meta-model combination, with model selection and weighting determined based on held-out validation set performance. These classifiers take in the 768-dimensional embeddings and carefully picked parts of them. This method makes it simpler to add a lot of neuroimaging features to a bigger multimodal prediction framework.

4.3.6 Multimodal Framework Integration

We carefully extracted and saved the 768-dimensional MRI embeddings for all the participants in the dataset, together with the predictions made by the trained unimodal classifiers. The embeddings give us a clear and useful picture of neuroanatomical structure, which helps us comprehend the behavioral patterns that show up in questionnaire responses (Module 2) and the demographic factors that give us important context at the population level (Module 3). Demographic variables also focus on social and developmental risk factors, whereas behavioral measurements give us important information about the psychological and symptomatic aspects of anxiety. There is a clear link between neuroimaging embeddings and observable changes in brain structure, which shows that they could reveal underlying neurobiological weaknesses.

The embeddings are used in the cross-modal attention-based fusion strategy that is explained in Section 4.8. This method needs trained attention processes that prioritize each sense based on how relevant it is to particular predictions in order to carefully combine input from all three senses. The framework may employ information that works well together because it has a lot of different types of information. When behavioral symptoms are ambiguous, distinctive neuroanatomical patterns may provide substantial predictive insights. Conversely, behavioral and demographic data can be quite beneficial when the brain's anatomy is somewhat similar.

This module elucidates the methodology of structural neuroimaging, examined through a deep learning framework specifically designed for the extraction of neuroanatomical features and initialized with pretrained weights derived from comprehensive brain age prediction datasets, yielding a significant predictive signal for the early identification of anxiety disorders. The precise spatial correlations within the resultant 768-dimensional space can proficiently depict complex structural characteristics of brain anatomy. These characteristics may reveal essential neurodevelopmental disparities influencing anxiety risk, which are frequently difficult to detect using traditional volume analysis, region-of-interest analysis, or normal radiologic interpretation. These measures can pick up on minor changes in the thickness of the cortex and the size of structures such as the amygdala in the brain.

The embedding provides a neurobiological basis for the multimodal prediction framework, providing a mechanistic basis that may be able to explain some of the behavioral/demographic risk factors identified within other modalities. The extracted embeddings are stored in a standard format together with their corresponding classifier prediction for maximal flexibility in integration approaches, ranging over a broad spectrum that starts at feature concatenation approaches to more complex feature integration through attention mechanisms within the multimodal framework. This provides maximal flexibility for adapting to different potential clinical application scenarios whilst ensuring that the full potential of structural brain imaging data is being used to produce optimal anxiety disorder prediction.

4.4 Module 2: Questionnaire Based Forecasting Module

4.4.1 Questionnaire Module Overview

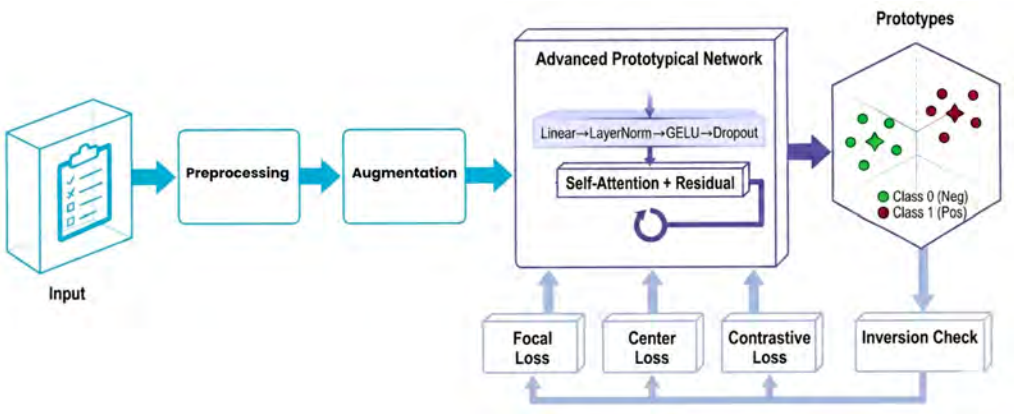


Figure 4.8: Questionnaire processing pipeline for Module 2.

Module 2 explores the examination of responses to anxiety-related questionnaires gathered during the initial assessment of the QTAB study. This module captures both self-reported and parent-reported symptoms across different areas of anxiety, offering valuable behavioral insights that complement neuroimaging findings and demographic data. The data acquired from this questionnaire corresponds with individual responses from standardized assessment tools employed to evaluate patients with anxiety disorders. This association indicates some patterns of symptoms, especially those that could mean someone has an anxiety problem. Module 2 employs a cautious feature selection strategy and a standard network architecture that is geared for few-shot learning when classification is not balanced. This happens because questionnaire data is sophisticated and there is a potential that data will leak from aggregate scores.
The module examines 109 initial parameters derived from established questionnaires for evaluating anxiety and mood, carefully eliminating aggregate scores to avert data leaking. We find and standardize the 40 most useful traits using correlation-based feature selection. A prototypical network architecture designed for few-shot learning processes the characteristics by embedding answers to surveys into a 32-dimensional space which is best for predicting anxiety. The architecture employs dual-encoder pathways with residual connections, self-attention mechanisms, and multi-head prototype learning to capture complex patterns in symptom expression. The 32-dimensional embeddings that come out from this process generate behavioral patterns that work with the neuroimaging and demographic modalities in the final multimodal fusion architecture.

4.4.2 Feature Extraction and Preprocessing

The features of the questionnaire were selected from the QTAB baseline assessment, resulting in an initial collection of 109 features based on standardized instruments for assessing anxiety and mood, which have been extensively validated in young children and adolescent groups. The characteristic set includes responses from the Spence Children's Anxiety Scale (SCAS) at the item level, along with the parent-reported version, known as the pSCAS. These tools assess six different dimensions of anxiety: separation anxiety, social phobic reactions, obsessive-compulsive traits, panic/agoraphobic conditions, fears related to physical injury, and general feelings of anxiousness. Additional items were sourced from instruments like the Short Mood and Feelings Questionnaire (SMFQ) and its parent-reported counterpart (pSMFQ), both of which assess depressive symptoms often linked to anxiety disorders. Furthermore, measures were taken from the Strengths and Difficulties Questionnaire (SDQ) and the parent-reported version (pSDQ), which evaluate broader emotional and behavioral challenges.

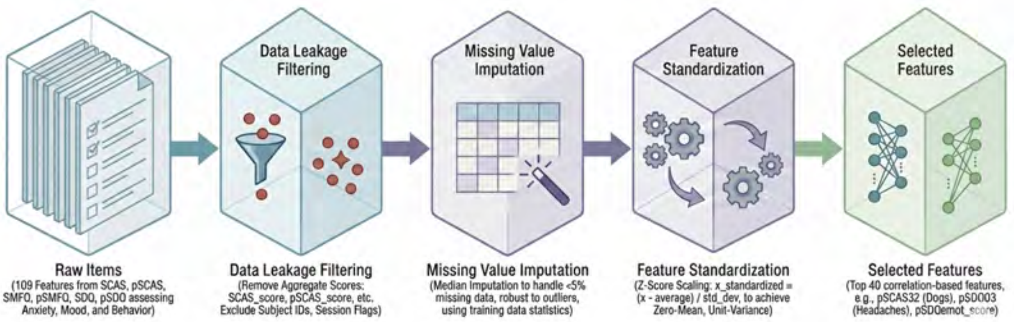


Figure 4.9: Questionnaire preprocessing pipeline showing data leakage filtering, missing value imputation, and feature standardization.

4.4.2.1 Data Leakage Prevention

In order to identify and get rid of aggregate scores that directly represent the intended outcome, a necessary preprocessing step was taken. This ensured that the prediction was possible and that simple solutions were avoided. The leaked features that were not included include: total anxiety scores shown by `SCAS_score`, `pSCAS_score`; the respective aggregate scores for domain-specific instruments shown by pairs like `SCAS_sepanx_score`, `pSCAS_sepanx_score`, `SCAS_panago_score`, `pSCAS_panago_score`; and depressive quotient columns denoted by `SMFQ_score`, `pSMFQ_score`. Additionally, non-feature columns related to subject identifiers (`subject_id`, `session`, substituted item flags, etc.) were excluded. This delicate filtering procedure creates 103 unique feature sets that can find certain patterns of anxiety-related symptom expression, even without direct diagnostic verification.

4.4.2.2 Missing Value Imputation

To fill in missing values in the dataset, we used median imputation, a reliable method that is less influenced by outliers and preserves the main data distributions unchanged. Notably, the median values were calculated on the original data before the actual data splitting into training-validation sets and a test set to avoid possible information leaks. This approach represents a practical compromise: while ideally imputation parameters would be estimated solely from training data, the small sample size and relatively low missing data rate ($< 5\%$ for most items) make training-only estimation unstable.

4.4.2.3 Feature Standardization

For data normalization, the feature matrix was standardized with the StandardScaler, which transforms features to have zero-mean and unit-variance distribution following the formula:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{4.2}$$

Equation (4.2) illustrates Z-score standardization, a method in which features are modified to have a mean of zero and a variance of one, as commonly discussed in data preprocessing literature [3].
where μ is the mean and σ is the standard deviation computed from the training set. This standardization is essential to incorporate all elements of the questionnaire equally in the training of the model without bias caused by features with the largest range.

4.4.3 Feature Selection Strategy

Given the high dimensionality (103 features) relative to the limited sample size characteristic of longitudinal neuroimaging cohorts, and recognizing the presence of redundant or weakly informative items across the multiple questionnaires, a correlation-based feature selection strategy was employed to identify the most discriminative symptom indicators.

4.4.3.1 Correlation-Based Ranking

The absolute Pearson correlation between each feature and the binary target outcome was computed on the combined training and validation sets. The features were organized based on the strength of their correlation, highlighting how strongly each one is associated with the presence of anxiety disorders. We selected the top 40 features, which account for approximately 39% of the original 103 features, for model training. This dimensionality represents a balance between maintaining sufficient information to understand the complex nature of anxiety symptoms and ensuring enough generalizability to prevent overfitting.

4.4.3.2 Selected Feature Characteristics

The ranges for these selected features were from an absolute high of 0.232 to approximately 0.15 minimum, suggesting fair but significant correlations with the anxiety outcome. Though these correlations seem low, when combined, they provide significant predictive value. The correlated features included items on parent-reported anxiety (e.g., pSCAS32, pSCAS19, pSCAS13), parent-reported emotions and behavioral factors (e.g., pSDQ03, pSDQemot_score, pSDQ24, pSDQ13), and aspects of anxiety that children reported using SCAS.

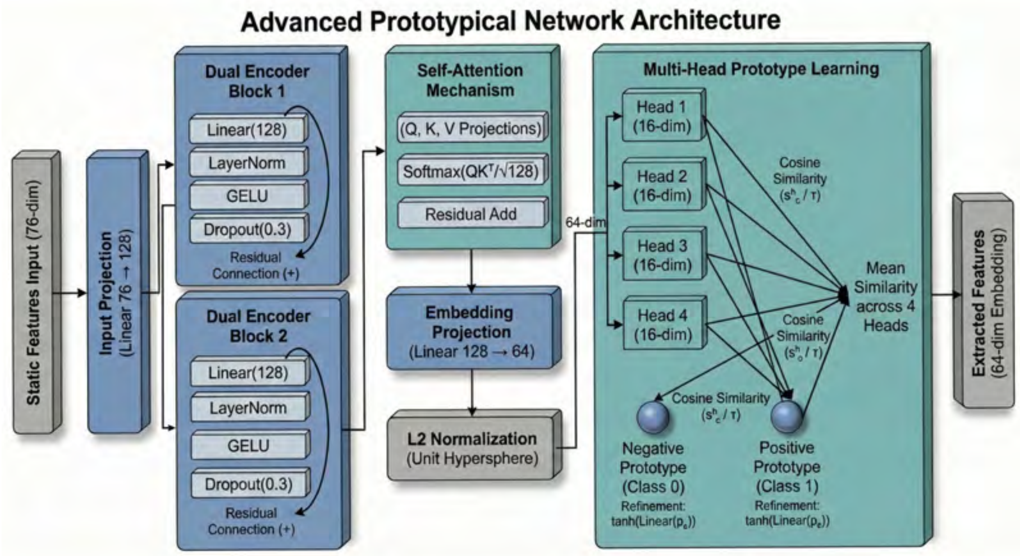


Figure 4.10: Detailed prototypical network architecture for questionnaire analysis.

4.4.4 Prototypical Network Architecture

The protocol followed in the questionnaires' analysis is a sophisticated prototypical network architecture designed for few-shot learning approaches in heavily high-dimensional data. Prototypical networks are designed after few-shot image classification approaches by Snell et al. (2017), in which classification is done in a metric space after the samples are embedded accordingly. This is highly appropriate for the anxiety prediction case since few samples are required due to the heavily class-imbalanced case.
The architecture is composed of four main components working in concert:

1. A feature projection component

2. A two-stream encoder structure with residual connection progression

3. A self-attention structure for handling feature complexities

4. A multi-head module for prototype computation

4.4.4.1 Input Projection Layer

The resulting 40-dimensional selected feature vector is first expanded to a higher-dimensional space of size 64 via a linear transformation:

$$\mathbf{h}_0 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \tag{4.3}$$

Equation (4.3) defines a linear input projection layer commonly used in neural networks to map low-dimensional feature representations into a higher-dimensional latent space [4].
where $\mathbf{x} \in R^{40}$ represents the input feature vectors after standardization. The expansion to a higher-dimensional space equips this network with greater representational ability to automatically apprehend more subtle non-linearities between the various items of the questionnaires.

4.4.4.2 Dual-Encoder Pathway

The projected features pass through two sequential encoder blocks, both incorporating a residual learning concept. Each encoder block consists of:

- A linear transformation maintaining dimensionality ($64 \to 64$)
- Layer Normalization to stabilize training
- GELU activation for smooth non-linearity
- Dropout with probability 0.2 to avoid overfitting

The first encoder block transforms the projected features:

$$\mathbf{h}_1 = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}\left(\text{Linear}(\mathbf{h}_0)\right)\right)\right) \tag{4.4}$$

The second encoder block processes $\mathbf{h}_1$:

$$\mathbf{h}_2 = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}\left(\text{Linear}(\mathbf{h}_1)\right)\right)\right) \tag{4.5}$$

The encoder structure in Equations (4.4)–(4.6), incorporating linear transformations, layer normalization, GELU activation, dropout, and residual connections, is inspired by Transformer-style encoder architectures [6].
Residual connections are made between encoder blocks:

$$\mathbf{h}_1^{\text{residual}} = \mathbf{h}_0 + \mathbf{h}_1 \tag{4.6}$$

$$\mathbf{h}_2^{\text{residual}} = \mathbf{h}_1 + \mathbf{h}_2 \tag{4.7}$$

These connections help permit gradient flow in backpropagation, arresting the vanishing gradient problem and allowing incremental feature refinements to be learned.

4.4.4.3 Self-Attention Mechanism

The resulting representation $\mathbf{h}_2^{\text{residual}}$ is input into a self-attention module. Query, key, and value projections are computed:

$$\mathbf{Q} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_Q \tag{4.8}$$

$$\mathbf{K} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_K \tag{4.9}$$

$$\mathbf{V} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_V \tag{4.10}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in R^{64\times 64}$ are weight matrices learned during training.
The attention weights are calculated through a softmax function:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{64}}\right) \tag{4.11}$$

where the scaling factor of $\sqrt{64}$ is used to avoid the "dot product explosion" phenomenon.
The attended output is calculated as:

$$\mathbf{h}_{\text{attended}} = \mathbf{A}\mathbf{V} \tag{4.12}$$

A residual connection yields:

$$\mathbf{h}_{\text{final}} = \mathbf{h}_2^{\text{residual}} + \mathbf{h}_{\text{attended}} \tag{4.13}$$

The attention mechanism defined in Equations (4.8)–(4.13), including query, key, and value projections with scaled dot-product attention and residual connections, is inspired by Transformer-based models [39].
This attention mechanism allows the model to weigh the interaction between features dynamically, enabling it to pick up on subtle relations between questionnaire responses not immediately apparent by simple linear correlation.

4.4.4.4 Embedding Projection

A final linear transformation integrates the features into a 32-dimensional embedding space:

$$\mathbf{e} = \text{Linear}(\mathbf{h}_{\text{final}}) \tag{4.14}$$

To make the embeddings the same length, L2 normalization is used:

$$\mathbf{e}_{\text{norm}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2} \tag{4.15}$$

All embeddings need to remain on the same hypersphere in a 32-dimensional space. This geometric restriction makes the embeddings good for calculating cosine similarity, which is the metric used in prototype-based categorization.

4.4.5 Multi-Head Prototype Learning

This proposed method uses a multi-head prototype learning strategy to make the embedding space stronger and more varied. Instead of showing the 32-dimensional representations as one big representation, they are separated into four heads, each with eight dimensions. This architecture allows various subspaces of the embedding to focus on identifying specific behavioral and psychological patterns linked to anxiety-related symptoms.

4.4.5.1 Prototype Initialization

For each class $c \in \{\text{anxiety-positive}, \text{anxiety-negative}\}$, a prototype is generated by computing the average of the embedding vectors from the support samples associated with that class:

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{i \in \mathcal{S}_c} \mathbf{e}_i \tag{4.16}$$

where $\mathcal{S}_c$ denotes the set of support samples from class c, N_c is the number of samples in that class, and $\mathbf{e}_i$ represents the embedding vector of the i-th sample. This mean prototype acts as the centroid of the class distribution in the embedding space.

4.4.5.2 Learnable Prototype Refinement

Instead of using the class mean as the final prototype, a learnable refinement mechanism is included. This improvement is made possible with a linear transformation followed by hyperbolic tangent activation:

$$\Delta\mathbf{p}_c = \tanh\left(\text{Linear}(\mathbf{p}_c)\right) \tag{4.17}$$

The linear layer takes the 32-dimensional prototype and transforms it into a refinement vector with the same dimensions. The refined prototype is calculated as:

$$\mathbf{p}_c^{\text{refined}} = \mathbf{p}_c + 0.1 \times \Delta\mathbf{p}_c \tag{4.18}$$

where the scaling factor of 0.1 limits the refinement to a small change, keeping the training stable while allowing for class-specific adjustments.
Finally, the refined prototypes are normalized to unit length:

$$\mathbf{p}_c^{\text{final}} = \frac{\mathbf{p}_c^{\text{refined}}}{\|\mathbf{p}_c^{\text{refined}}\|_2} \tag{4.19}$$

4.4.5.3 Multi-Head Similarity Computation

During inference, similarity between query embeddings and class prototypes is computed independently across the four heads. The embedding vector $\mathbf{e} \in R^{32}$ is reshaped into four subvectors:

$$\{\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3, \mathbf{e}^4\}, \quad \mathbf{e}^h \in R^8 \tag{4.20}$$

Each class prototype is decomposed similarly into $\{\mathbf{p}_c^h\}$. For each head h and class c, cosine similarity is computed as:

$$s_c^h = \frac{\mathbf{e}^h \cdot \mathbf{p}_c^h}{\|\mathbf{e}^h\|_2 \|\mathbf{p}_c^h\|_2} \tag{4.21}$$

Since all embeddings and prototypes are L2-normalized, the similarity reduces to a simple dot product. A learnable temperature parameter τ, initialized at 0.5 and clamped to a minimum value of 0.1 during training, scales the similarity scores:

$$\tilde{s}_c^h = \frac{s_c^h}{\tau} \tag{4.22}$$

Lower values of τ yield sharper confidence distributions, while higher values produce softer predictions. The final similarity score for each class is obtained by averaging across heads:

$$s_c = \frac{1}{4} \sum_{h=1}^{4} \tilde{s}_c^h \tag{4.23}$$

Through the use of softmax function class probabilities are then calculated :

$$P(c \mid \mathbf{e}) = \frac{\exp(s_c)}{\sum_{c'} \exp(s_{c'})} \tag{4.24}$$

The prototype computation and learnable refinement in Equations (4.16)–(4.19) follow standard prototypical network strategies, where class centroids are computed from support samples and optionally refined via a learnable transformation. Equations (4.20)–(4.24) implement the multi-head prototype similarity computation, where embeddings and prototypes are divided into subspaces, cosine similarity is computed per head, and a learnable temperature scales the scores before softmax-based probability estimation, following standard practices in prototypical network and metric-based few-shot learning frameworks [11].
This multi-headed approach works like an implicit ensemble, allowing various brains to pick up on various patterns of behavior that work well together while making predictions more stable and reliable.

4.4.6 Loss Function Design

The training goal consists of three important loss components that work together to resolve class imbalance, promote compact embedding structures, and make sure that classes are clearly separated.

4.4.6.1 Focal Loss

The primary classification objective is focal loss, which extends the standard cross-entropy formulation to emphasize difficult and minority-class samples:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha(1-\hat{p}_i)^{\gamma}\log(\hat{p}_i) \tag{4.25}$$

where $\hat{p}_i$ denotes the predicted probability of the true class, $\alpha = 0.75$ balances class importance, and $\gamma = 2.0$ controls the focusing strength. This loss minimizes the impact of easily categorized negative samples and shifts learning toward difficult and minority-class examples.

4.4.6.2 Center Loss

A center loss phrase is added to encourage compactness within a class:

$$\mathcal{L}_{\text{center}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{e}_i - \mathbf{c}_{y_i}\|_2^2 \tag{4.26}$$

where $\mathbf{c}_{y_i}$ is the learnable center corresponding to class label y_i. The class centers are optimized jointly with model parameters. A small weighting factor of 0.01 ensures that this loss guides cluster formation without dominating the primary objective.

4.4.6.3 Contrastive Loss

A margin-based contrastive loss is applied exclusively to positive samples to reinforce clear decision boundaries:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{N_{\text{pos}}}\sum_{i\in\text{pos}} \max\left(0,\; m - \cos(\mathbf{e}_i, \mathbf{p}_{\text{pos}}) + \cos(\mathbf{e}_i, \mathbf{p}_{\text{neg}})\right) \tag{4.27}$$

where $m = 0.3$ defines the required separation margin. This term penalizes embeddings that lie too close to the negative prototype, explicitly enforcing discriminative separation in the low-data regime.

4.4.6.4 Total Objective

The complete training objective is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + 0.01 \times \mathcal{L}_{\text{center}} + 0.3 \times \mathcal{L}_{\text{contrastive}} \tag{4.28}$$

The training objective in Equations (4.25)–(4.28) combines focal loss for class-imbalanced classification [9], center loss to encourage intra-class compactness [8], and contrastive loss to enforce inter-class separation [5], following established practices in deep learning [6].
Gradient clipping with a maximum L2 norm of 1.0 is applied to stabilize optimization when combining losses of different magnitudes.

4.4.7 Data Augmentation and Training Procedure

To fix the problem of a very uneven class distribution, where only 13.3% of samples belong to the positive class, a copy-paste augmentation approach is used during training. Each positive example is copied and Gaussian noise is added:

$$\mathbf{x}_{\text{aug}} = \mathbf{x}_{\text{original}} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, 0.05^2\mathbf{I}) \tag{4.29}$$

The copy-paste Gaussian noise augmentation in Equation (4.29) follows standard data augmentation practices used to mitigate class imbalance in deep learning [12]. This represents a 5% change after standardization, which doubles the number of positive samples from 31 to 62. Importantly, augmentation is only applied to the training set; the validation and test sets remain unaugmented for fair evaluation.

4.4.7.1 Episodic Training Strategy

Model training uses an episodic learning paradigm based on few-shot learning frameworks. A 60:40 stratified split is used to randomly separate the augmented training data into support and query sets at each epoch. The support set is used to compute prototypes, which are then used to classify query samples using the multi-head similarity approach. This episodic approach promotes the model to frequently acquire generalizable prototype representations.

4.4.7.2 Optimization and Model Selection

The AdamW optimizer trains the network for 150 epochs, with:

- Initial learning rate: 0.002
- Weight decay: 0.01
- One-cycle learning rate schedule (peak at 0.005 during first 30% of training)

The validation AUC-ROC is used for model selection. The checkpoint with the greatest validation AUC is kept for final test evaluation.

4.4.8 Feature Embedding Extraction

For multimodal fusion, the trained prototypical network serves as a feature extractor, producing 32-dimensional embeddings for each participant that encapsulate their questionnaire response patterns in a compact, discriminative space optimized for anxiety disorder prediction. These embeddings are produced by executing the series of operations (input projection, dual encoder with residual connections, self-attention layer, final projection layer) onto the 40 standardized features, followed by L2 normalization.

The extraction process is performed in inference mode (with dropout disabled) to ensure deterministic, reproducible embeddings. These extracted embeddings capture learned "signatures" relating to symptom profiles, informant concurrence or non-concurrence, and intricate configurations of anxiety manifestations.

The model achieved an AUC on the test set of 0.782, indicating that questionnaire data, when properly preprocessed to avoid data leakage, provides strong predictive information for anxiety disorder identification. The 32-dimensional embedding space offers a robust, precise representation that enhances the neuroimaging as well as static demographic modalities inside the ultimate multimodal prediction system.

4.5 Module 3: Phenotypic- Based Forecasting Module

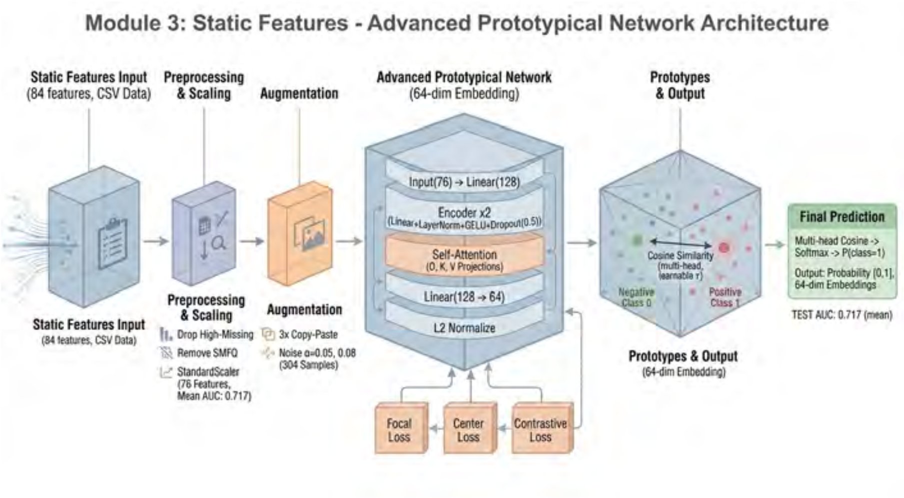


Figure 4.11: Demographic/static feature pipeline producing 64-dimensional embeddings.

Module 3 emphasizes the integration of static phenotypic characteristics, encompassing demographic variables, socioeconomic indicators, and foundational physiological measures, to augment neuroimaging and behavioral methodologies. Unlike the neuroanatomical patterns described in Module 1 and the clinical symptoms detailed in Module 2, static features function as persistent, time-invariant risk indicators that reflect intrinsic predispositions and environmental factors associated with adolescent anxiety disorders. These traits show how age, sex, socioeconomic level, and baseline physiological condition can all affect the environment where anxiety risk grows. The module employs a prototypical network architecture designed to address the challenges associated with static feature modeling, including weak individual feature correlations (maximum absolute correlation: 0.187), high-dimensional constrained representations (76 features from an initial 84), and significant class imbalance (13.3% positive class prevalence).

4.5.1 Feature Extraction and Preprocessing

Static features from the QTAB preliminary exam have been used to build an initial list of 84 features from diverse categories that represent each participant's stable attributes and environmental background. The full list of features includes:
Socio Demographic Features: Fundamental demographic variables encompass sex (binary), age at baseline assessment (continuous), and family structure indicators that reflect household composition and parental connections.

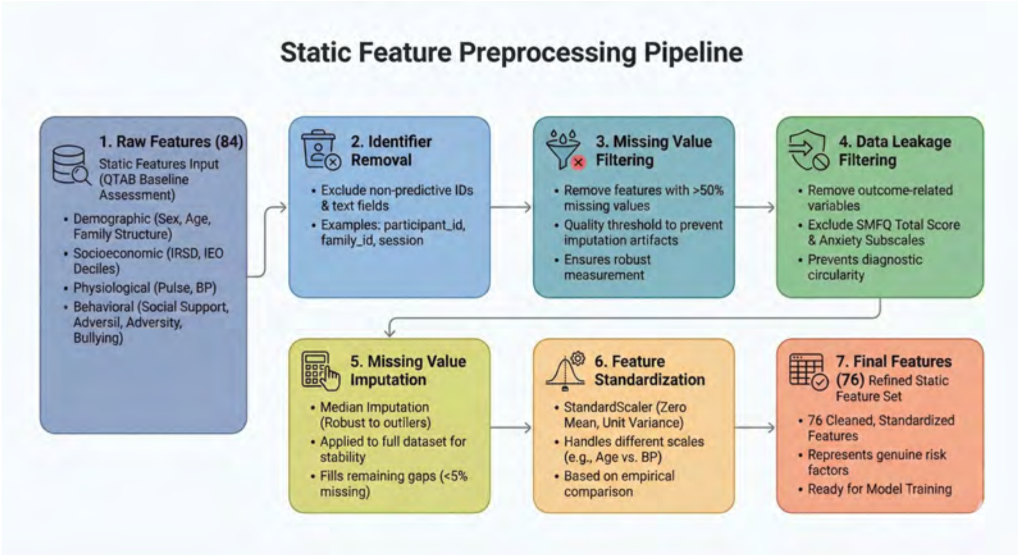


Figure 4.12: Static feature preprocessing pipeline for Module 3.

Social and Economic Indicators: The Index of Relative Socioeconomic Disadvantage (IRSD) along with the Index of Education and Occupation (IEO) are two socioeconomic indicators that are particular to certain areas and were based on Australian census data and the participants' home postcodes. The information was displayed in deciles. These indices indicate how a person's mental health is related to their socioeconomic background at the population level. They indicate that kids are more likely to get anxious and other mental diseases when they have challenges and restricted opportunities of attending school or work.

Physiological Measures: During the first visit to the doctor, physiological measurements were collected, including steady heart rate (beats per minute), pulmonary blood pressure (mmHg), and diastolic blood pressure (mmHg). These autonomic assessments may reveal underlying patterns of physiological arousal or stress response that could be associated with an increased propensity to perceive concern.

Initial Behavioral Measures: Social support scales (which measure how much support a person feels from family, friends, and important other people), family adversity scores (which measure how much family conflict, parental mental health problems, and dysfunctional households a person is exposed to), as well as bullying victimization dices (which measure how often a person has been bullied by a peer) are all examples of psychosocial context variables. These environmental factors illustrate stressful circumstances and protective elements that alter the risk of anxiety via psychological mechanisms.

The QTAB dataset include twin pairs, which makes it possible to study the genetic and environmental factors that affect anxiety risk. There are 422 people from 211 homes, and 222 of them are monozygotic (MZ) twins (52.6

Twin twins serve as a natural check for common genetic and environmental influences. Does the integration of twin data (monozygotic and dizygotic pairs) with T1-weighted MRI, anxiety questionnaires, and demographic information enhance the early prediction of anxiety in adolescents from baseline to the initial follow-up, in comparison to ABCD? This structure enables the model to discern patterns that differentiate genetic predisposition from environmental risk factors, potentially enhancing prediction through the utilization of within-pair similarities and differences.

The final set of 76 characteristics includes zygosity indicators (zyg_MZ, zyg_DZ) as static features that capture twin-specific genetic relationships. Although family_id identifiers were excluded from the feature set during preprocessing as administrative variables, they were used to organize the dataset and preserve familial structure.
By incorporating zygosity information, the prototypical network can learn repre sentations that reflect patterns of genetic similarity. For instance, the model might show that monozygotic (MZ) twins with similar demographic, socioeconomic, and physiological characteristics had more similar anxiety-related outcomes than dizygotic (DZ) twins with comparable attributes. These results suggest a more significant genetic impact in MZ twin pairings.
The attention process also helps the model figure out how zygosity and other fixed qualities affect each other. This suggests that some combinations of risk factors may more accurately forecast outcomes in MZ pairs than in DZ couples. This version lacks clear genetic markers, such as polygenic risk scores, candidate gene variants, or epigenetic markers. However the twin-based structure uses zygosity as a stand-in for genetic susceptibility.

4.5.1.1 Data Leakage Prevention and Quality Control

In order to prevent data leakage, which happens when information about the result mistakenly influences the forecast, a tight preparation pipeline with various filtering steps was put in place:
First, the feature set was cleaned up by removing columns that identify people (participant_id, family_id, session) and text fields that don't have numbers or give any forecast information. We utilize these administrative criteria to keep track of data, but they fail to inform us anything about the risk of anxiousness.
Second, features with more than half of their values missing were removed so that errors in imputation wouldn't ruin the image. When more than half of the values for an element are missing, the values that are estimated from the ones that are left may not show the true distribution and could cause false patterns. This quality level ensures that the study's features can be accurately measured by most participants.
Third, features directly related to the target outcome were identified and excluded as leaky variables that would enable trivial prediction through diagnostic circularity. Specifically, the Short Mood and Feelings Questionnaire (SMFQ) total score was removed, as this aggregate depression measure directly quantifies mood symptomatology that substantially overlaps with anxiety disorders and could enable the model to achieve high accuracy by simply recognizing comorbid depression rather than predicting anxiety risk from independent risk factors.
This filtering procedure resulted in a final feature set of 76 static variables that represent genuine risk factors and contextual characteristics, free from information leakage and measurement quality concerns.

4.5.1.2 Missing Value Imputation

Missing value imputation was performed using median imputation a robust strategy that replaces missing values with the median of the observed values for each feature. The median is preferred over the mean for its robustness to outliers and its appropriateness for both continuous and ordinal features.

4.5.1.3 Feature Standardization

The final feature matrix was standardized using StandardScaler, which transforms each feature to have zero mean and unit variance through the transformation:

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma} \tag{4.30}$$

The feature matrix standardization in Equation (4.30) uses zero-mean, unit-variance scaling (Z-score normalization) following standard practices in machine learning [3]. where μ and σ are the mean and standard deviation computed from the training set. This standardization is essential because static features span widely different scales—age ranges from approximately 9 to 16 years, blood pressure ranges from 50 to 150 mmHg, IRSD deciles range from 1 to 10—and without standardization, features with larger numerical ranges would dominate the learning process.

4.5.2 Prototypical Network Architecture

Given the weak signal-to-noise ratio in static features evidenced by the maximum absolute correlation with the anxiety outcome being only 0.187, indicating that no single static feature exhibits strong univariate association with the target and the limited sample size, a prototypical network architecture was selected for its ability to learn robust class representations through few-shot learning principles.

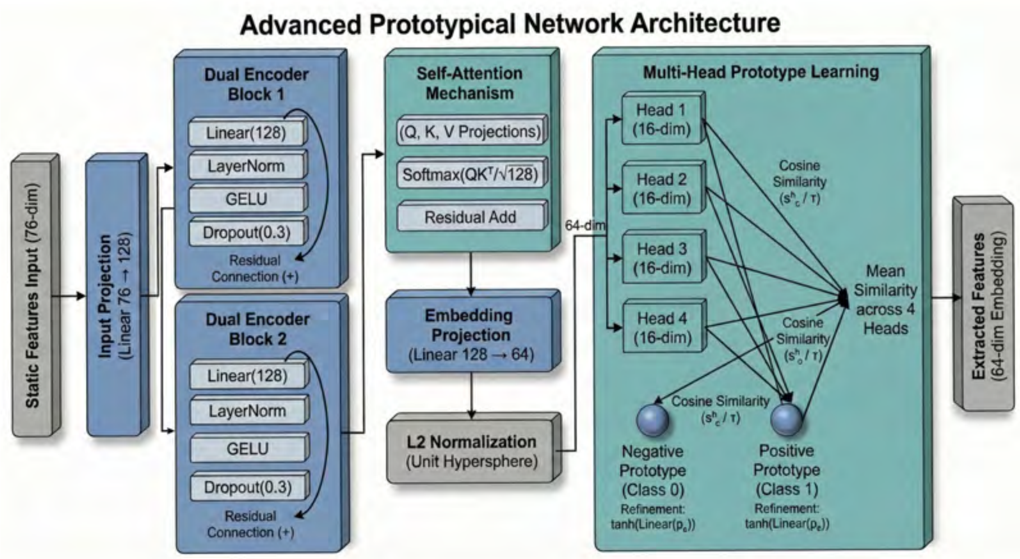


Figure 4.13: Advanced prototypical network architecture for static features with multi-head prototype learning.

The network architecture consists of four primary components working in concert:

1. A feature projection layer that expands the input dimensionality to provide representational capacity
2. A dual-encoder pathway with residual connections that progressively transforms features into increasingly abstract representations

3. A self-attention mechanism that captures complex dependencies between static features

4. A multi-head prototype computation module that learns robust class representations

4.5.2.1 Input Projection Layer

The 76-dimensional static feature vector is projected into a 128-dimensional hidden space via a linear transformation:

$$\mathbf{h}_0 = \mathbf{W}_1\mathbf{x} + \mathbf{b}_1 \tag{4.31}$$

where $\mathbf{x} \in R^{76}$ represents the standardized input features and $\mathbf{h}_0 \in R^{128}$ represents the projected hidden state. This expansion into a higher-dimensional space (approximately 1.7× expansion),it might more effectively characterize complicated, non-linear interconnections across demographic, socioeconomic, physiological, and psychosocial variables.

4.5.2.2 Dual-Encoder Pathway

The projected features pass through two encoder blocks in a row, and each one utilizes a residual learning scheme. Each encoder block comprises: (1) a linear transformation mapping the 128-dimensional input to 128-dimensional output, (2) layer normalization, (3) GELU (Gaussian Error Linear Unit) activation, and (4) dropout with probability 0.3.
The first encoder block transforms the projected features:

$$\mathbf{h}_1 = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}\left(\text{Linear}(\mathbf{h}_0)\right)\right)\right) \tag{4.32}$$

The second encoder block processes $\mathbf{h}_1$ similarly:

$$\mathbf{h}_2 = \text{Dropout}\left(\text{GELU}\left(\text{LayerNorm}\left(\text{Linear}(\mathbf{h}_1)\right)\right)\right) \tag{4.33}$$

Residual connections combine the input and output of each block:

$$\mathbf{h}_1^{\text{residual}} = \mathbf{h}_0 + \mathbf{h}_1 \tag{4.34}$$

$$\mathbf{h}_2^{\text{residual}} = \mathbf{h}_1 + \mathbf{h}_2 \tag{4.35}$$

4.5.2.3 Self-Attention Mechanism

The encoded representation $\mathbf{h}_2^{\text{residual}}$ is passed through a self-attention module. Three linear transformations produce query, key, and value projections:

$$\mathbf{Q} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_Q \tag{4.36}$$

$$\mathbf{K} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_K \tag{4.37}$$

$$\mathbf{V} = \mathbf{h}_2^{\text{residual}}\mathbf{W}_V \tag{4.38}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in R^{128\times128}$ are learned projection matrices.

Attention weights are computed as:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{128}}\right) \tag{4.39}$$

The attention output $\mathbf{h}_{\text{attended}} = \mathbf{AV}$ is combined with the original representation via residual connection:

$$\mathbf{h}_{\text{final}} = \mathbf{h}_2^{\text{residual}} + \mathbf{h}_{\text{attended}} \tag{4.40}$$

4.5.2.4 Embedding Projection

The attended features are projected into a 64-dimensional embedding space through a final linear transformation:

$$\mathbf{e} = \text{Linear}(\mathbf{h}_{\text{final}}) \tag{4.41}$$

L2 normalization is applied to ensure unit-length embeddings:

$$\mathbf{e}_{\text{norm}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2} \tag{4.42}$$

4.5.3 Multi-Head Prototype Learning

For prototype analysis, a multi-head design is used. This makes it more dependable and enables different parts of the embedding focus on capturing distinctive static indicators of risk. The 64-dimensional embedding is conceptually partitioned into four 16-dimensional heads.

4.5.3.1 Prototype Initialization

For each class $c \in \{\text{anxiety-positive}, \text{anxiety-negative}\}$, a prototype is computed as the mean embedding of all support samples belonging to that class:

$$\mathbf{p}_c = \frac{1}{N_c} \sum_{i \in \mathcal{S}_c} \mathbf{e}_i \tag{4.43}$$

where $\mathcal{S}_c$ represents the set of support samples from class c, N_c is the number of samples in that class, and $\mathbf{e}_i$ is the embedding of sample i.

4.5.3.2 Learnable Prototype Refinement

A learnable refinement step adaptively adjusts prototype locations. The refinement module consists of a linear transformation followed by hyperbolic tangent activation:

$$\Delta\mathbf{p}_c = \tanh\left(\text{Linear}(\mathbf{p}_c)\right) \tag{4.44}$$

The refined prototype is computed as:

$$\mathbf{p}_c^{\text{refined}} = \mathbf{p}_c + 0.1 \times \Delta\mathbf{p}_c \tag{4.45}$$

The refined prototypes are L2-normalized to maintain unit length:

$$\mathbf{p}_c^{\text{final}} = \frac{\mathbf{p}_c^{\text{refined}}}{\|\mathbf{p}_c^{\text{refined}}\|_2} \tag{4.46}$$

4.5.3.3 Multi-Head Similarity Computation

The input embedding $\mathbf{e} \in R^{64}$ is reshaped into four 16-dimensional head representations: $\{\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^3, \mathbf{e}^4\}$. For each head h and each class c, cosine similarity is computed:

$$s_c^h = \frac{\mathbf{e}^h \cdot \mathbf{p}_c^h}{\|\mathbf{e}^h\|_2 \|\mathbf{p}_c^h\|_2} \tag{4.47}$$

A learnable temperature parameter τ (initialized at 0.5, clamped to a minimum of 0.1) scales the similarity scores:

$$s_{c,\text{scaled}}^h = \frac{s_c^h}{\tau} \tag{4.48}$$

For every class, the mean similarity across heads is calculated as below:

$$s_c = \frac{1}{4} \sum_{h=1}^{4} s_{c,\text{scaled}}^h \tag{4.49}$$

Lastly, the softmax normalization gives us the final probability for each class:

$$P(c \mid \mathbf{e}) = \frac{\exp(s_c)}{\sum_{c'} \exp(s_{c'})} \tag{4.50}$$

4.5.4 Loss Function Design

The training objectives include three loss concepts which function together to fix class imbalance, encourage embeddings that can tell the difference between classes, and promote class prototypes that are separated effectively.

4.5.4.1 Focal Loss

The primary loss component is focal loss with $\alpha = 0.75$ and $\gamma = 2.0$:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N} \sum_{i=1}^{N} \begin{cases} \alpha(1-\hat{p}_i)^\gamma \log(\hat{p}_i) & \text{if } y_i = 1 \\ (1-\alpha)\hat{p}_i^\gamma \log(1-\hat{p}_i) & \text{if } y_i = 0 \end{cases} \tag{4.51}$$

where $\hat{p}_i$ is the predicted probability for the positive class.

4.5.4.2 Center Loss

A center loss term with weight 0.01 encourages embeddings to cluster around learnable class centers:

$$\mathcal{L}_{\text{center}} = \frac{0.01}{N} \sum_{i=1}^{N} \|\mathbf{e}_i - \mathbf{c}_{y_i}\|_2^2 \tag{4.52}$$

where $\mathbf{e}_i$ is the embedding of sample i, y_i is its true class label, and $\mathbf{c}_{y_i}$ is the learnable center for that class.

4.5.4.3 Contrastive Loss

A contrastive loss term with weight 0.3 explicitly enforces separation between positive samples and the negative prototype:

$$\mathcal{L}_{\text{contrastive}} = \frac{0.3}{N_{\text{pos}}} \sum_{i \in \text{positive}} \max\left(0,\ 0.3 - \cos(\mathbf{e}_i, \mathbf{p}_{\text{pos}}) + \cos(\mathbf{e}_i, \mathbf{p}_{\text{neg}})\right) \tag{4.53}$$

4.5.4.4 Total Loss

The complete training objective combines all three components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{focal}} + 0.01 \times \mathcal{L}_{\text{center}} + 0.3 \times \mathcal{L}_{\text{contrastive}} \tag{4.54}$$

Gradient clipping with maximum L2 norm of 1.0 is applied during backpropagation to prevent exploding gradients.

4.5.5 Data Augmentation and Training Procedure

To mitigate the effects of severe class imbalance (13.3% positive class prevalence, corresponding to 31 positive samples out of 242 total in the training set), a copy-paste augmentation strategy was employed.

4.5.5.1 Copy-Paste Augmentation

Each positive class sample in the training set was duplicated twice, with each duplicated copy augmented by adding Gaussian noise:

$$\mathbf{x}_{\text{aug1}} = \mathbf{x}_{\text{original}} + \boldsymbol{\epsilon}_1, \quad \boldsymbol{\epsilon}_1 \sim \mathcal{N}(\mathbf{0}, 0.05^2\mathbf{I}) \tag{4.55}$$

$$\mathbf{x}_{\text{aug2}} = \mathbf{x}_{\text{original}} + \boldsymbol{\epsilon}_2, \quad \boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, 0.08^2\mathbf{I}) \tag{4.56}$$

This augmentation effectively increases the number of positive samples from 31 to 93, shifting the class distribution from 12.8% (31/242) to approximately 30.6% (93/304) positive class prevalence.

4.5.5.2 Episodic Few-Shot Learning Paradigm

Training follows an episodic learning strategy where, at each training epoch, the augmented training dataset is randomly partitioned into support (60%, approximately 182 samples) and query (40%, approximately 122 samples) sets with stratification to maintain class balance. Prototypes are computed from the support set, and predictions are made on the query set by computing multi-head similarities between query embeddings and the support-derived prototypes.

4.5.5.3 Optimization and Regularization

The model is trained for 250 epochs using the AdamW optimizer with initial learning rate 0.001 and weight decay 0.02. A one-cycle learning rate schedule is employed, with the learning rate increasing linearly from 0.001 to a maximum of 0.003 over the first 30% of training epochs (epochs 0–75), then decaying back to zero over the remaining 70% of training (epochs 75–250).

4.5.5.4 Model Selection

Model performance is monitored on the validation set after each epoch, with the area under the receiver operating characteristic curve (AUC-ROC) serving as the primary evaluation metric. The model checkpoint corresponding to the best validation AUC across all 250 epochs is saved and used for final test set evaluation.

4.5.6 Feature Embedding Extraction and Performance

For multimodal fusion (described in Section 4.8), the trained prototypical network serves as a feature extractor, producing 64-dimensional embeddings for each participant that encapsulate their static feature constellation in a compact, discriminative space optimized for anxiety disorder prediction. These embeddings are extracted by passing the standardized 76-dimensional static features through the complete network input projection, dual-encoder pathway with residual connections, self-attention mechanism, and final embedding projection and collecting the L2-normalized output from the 64-dimensional projection layer, prior to prototype computation and classification.

The extraction process is performed in inference mode (with dropout disabled) to ensure deterministic, reproducible embeddings. For all participants across training, validation, and test sets, the same trained network weights are used to extract embeddings, ensuring consistency. These 64-dimensional embeddings are subsequently saved alongside the embeddings from the MRI Modality Based Forecasting Module (768-dimensional) and the questionnaire module (32-dimensional) for integration in the cross-modal attention-based fusion framework.

The embeddings capture learned demographic and contextual signatures that reflect stable risk factors socioeconomic disadvantage, physiological arousal patterns, psychosocial adversity in a form optimized for anxiety prediction. By operating in a learned metric space, these embeddings enable sophisticated similarity-based reasoning about how the combination of demographic, socioeconomic, physiological, and psychosocial factors positions an individual relative to the typical anxiety-positive versus anxiety-negative profiles.

Summary of the Performance: The static feature module achieves a test set AUC of 0.717 (mean across multiple runs: 0.717 ± 0.05), indicating that static features provide significant predictive signal when utilized within an architecture optimized for few-shot learning and resilient representation extraction in low signal-to-noise environments, despite their individual correlations being weak (maximum absolute correlation: 0.187). This finding corroborates the notion that anxiety risk originates from intricate multivariate patterns in demographic and environmental characteristics, rather than from robust individual risk indicators, and that deep learning architectures are capable of identifying these pattern-recognition tools.

The 64-dimensional embedding space provides a stable, well-calibrated representation that complements the neuroimaging and behavioral modalities in the final multimodal prediction framework. Where Module 1 represents neurobiological substrates, Module 2 represents current symptomatic expression, and Module 3 represents stable background risk and context information. This synergistic and complementary relationship between the three modules provides the foundation for decision-level fusion and combines the synergistic benefits of utilizing the information contained within each of these modalities—neuroimaging providing brain-based prediction of vulnerability, subjective reporting of behavior providing expression of illness, and static features providing context information for the expression of anxiety disorders. These three perspectives are combined and form the foundation of the comprehensive multimodal prediction and decision algorithm.

4.6 Multimodal Fusion Architecture

The multimodal fusion architecture is the final component in the entire framework. Here, predictions drawn from each of three separately prepared data encoders—structural MRI information, responses to questionnaires, and static demographic data—are fused to arrive at a unified and complete evaluation of risk with regard to anxiety disorder. The problem of accurately predicting risk in mental health disorders is a major concern in data modeling within health disorders because any form of data on its own cannot comprehensively cover the complexity of a mental health concern. Disorders in mental health result from a mixture of factors such as interactions between the brain, behavioral patterns, and environmental factors.
Because neuroimaging datasets typically have limited sample sizes and each modality provides complementary information, a decision-level fusion strategy is used, where well-calibrated probability predictions from each module are combined through weighted averaging. This approach differs from feature-level fusion, which would combine embeddings from all three modules and learn cross-modal interactions using additional neural network layers. Decision-level fusion offers several advantages in this context:

- It preserves the discriminative representations learned by each modality-specific encoder without requiring joint retraining
- It remains stable under limited data conditions by avoiding complex high-dimensional interaction learning that could lead to overfitting
- It allows interpretable weighting of each modality's contribution to the final prediction, supporting clinical understanding and model transparency

The architecture of the fusion uses weighted late fusion, whereby the probabilities from each modality are combined in a weighted linear sum. For each individual i, the final multimodal prediction is computed as:

$$P_{\text{multimodal}}(i) = w_1 \times P_{\text{MRI}}(i) + w_2 \times P_{\text{Questionnaire}}(i) + w_3 \times P_{\text{Static}}(i) \tag{4.57}$$

where $P_{\text{MRI}}(i)$, $P_{\text{Questionnaire}}(i)$, and $P_{\text{Static}}(i)$ denote the probability prediction outcomes of the first, second, and third modules, respectively, for the i-th participant. The weights (w_1, w_2, w_3) are positive values that satisfy the unity constraint:

$$w_1 + w_2 + w_3 = 1 \tag{4.58}$$

The final multimodal risk probability computation using weighted decision-level fusion is defined in Equations (4.57) and (4.58), subject to a unity constraint on the fusion weights [52].
which enforces the satisfaction of the probability within the range $[0, 1]$ and interprets $P_{\text{multimodal}}$ as a calibrated risk probability, quantifying the model's prediction for the i-th subject to develop an anxiety disorder.

4.6.1 Rationale for Decision-Level Fusion

The decision-level fusion technique is selected as opposed to other possibilities, namely feature-level fusion (where embeddings are concatenated followed by classification) and hybrid fusion (where both feature-level and decision-level fusion are used). The rationale for this selection is as follows:

First, the encoders for each modality have already been fine-tuned individually based on the respective input type through extensive end-to-end training with modality-specific objectives. The MRI-based encoder has undergone pretraining in brain age prediction and fine-tuning for anxiety classification. Similarly, the questionnaire and static feature modules undergo extensive end-to-end fine-tuning with focal, center, and contrastive losses to address class imbalance and weak signals.

Second, learning feature-level fusion interactions between different modalities from only 242 training samples presents significant challenges. The combined embedding would have $768 + 32 + 64 = 864$ dimensions:

$$\mathbf{e}_{\text{combined}} \in R^{864} = R^{768}_{\text{MRI}} \oplus R^{32}_{\text{Quest}} \oplus R^{64}_{\text{Static}} \tag{4.59}$$

Feature-level fusion, as expressed in Equation (4.59), would require concatenating modality-specific embeddings into a high-dimensional representation, substantially increasing model complexity [24].

Learning prediction functions from such a high-dimensional feature space with limited samples creates high risk of overfitting. For decision-level fusion, only three probability values must be combined, making it more stable and reliable.

Third, late fusion provides clear interpretability because the model's coefficients $\{w_1, w_2, w_3\}$ clearly present the contribution level each modality provides to the final prediction. For example, if $w_2 = 0.63$, it is clear that questionnaire responses contributed 63% to the final decision.

Fourth, decision-level fusion allows system modularity, as the system can function even when certain modalities are unavailable for certain patients by adjusting the weights accordingly.

4.6.2 Weight Optimization Strategy

Since the three modalities provide complementary information and their contributions must be carefully balanced, the optimal set of fusion weights is found using a structured grid search approach. The objective function to be maximized is predictive performance on unseen data.

4.6.2.1 Search Space Definition

The weight search space is defined using the following constraints:

$$w_1 \in [0.15, 0.45] \quad \text{(MRI weight)} \tag{4.60}$$

$$w_2 \in [0.35, 0.65] \quad \text{(Questionnaire weight)} \tag{4.61}$$

$$w_3 \in [0.10, 0.45] \quad \text{(Static feature weight)} \tag{4.62}$$

with a step size of 0.02 to allow fine-grained exploration, subject to the normalization constraint:

$$w_1 + w_2 + w_3 = 1 \tag{4.63}$$

The constrained fusion weight optimization strategy is formalized in Equations (4.60)–(4.63) [52].
These ranges are motivated by unimodal performance:

- Questionnaire Based Forecasting Module: strongest individual performance (AUC = 0.78), requiring significant weighting (35%–65%)
- MRI Modality Based Forecasting Module: intermediate performance (AUC = 0.75), requiring significant but not dominant weighting (15%–45%)
- Phenotypic-Based Forecasting Module: lowest individual performance, requiring substantial but subordinate weighting (10%–45%)

4.6.2.2 Grid Search Procedure

The validation set is used to optimize the weights so that the model doesn't fit too closely to the test set. This is how the grid search works:

1. **Generate candidate weight configurations:** Enumerate all weight combinations $\{w_1, w_2, w_3\}$ where each weight falls within its specified range with step size 0.02, and the weights sum to 1.0 (within numerical tolerance $\epsilon = 0.001$). This generates approximately 200–300 valid weight combinations.

2. **Compute fused predictions:** For each candidate weight configuration, compute the fused probabilities for all validation set participants:

$$P_{\text{multimodal}} = w_1 \times P_{\text{M1}} + w_2 \times P_{\text{M2}} + w_3 \times P_{\text{M3}} \tag{4.64}$$

   For each candidate weight configuration, fused multimodal predictions are computed using Equation (4.64) [38].

3. **Evaluate performance:** Compute the area under the ROC curve (AUC-ROC) for the fused predictions on the validation set.

4. **Select optimal weights:** Identify the weight configuration $\{w_1^*, w_2^*, w_3^*\}$ that achieves the maximum validation set AUC:

$$\{w_1^*, w_2^*, w_3^*\} =_{w_1,w_2,w_3} \text{AUC}_{\text{validation}}(w_1, w_2, w_3) \tag{4.65}$$

   The optimal fusion weights are selected by maximizing validation-set AUC as formalized in Equation (4.65) [38].

5. **Performance evaluation:** Log both validation set performance (used for model selection) and test set performance for post-hoc analysis.

This method of choosing weights based on validation makes sure that they work for data that has not been seen previously. Two other more advanced optimization approaches that could be employed are Bayesian optimization and gradient-based algorithms. But exhaustive grid search offers a lot of advantages: it doesn't need any more hyperparameters, it reveals the complete optimization landscape, and you can conduct it on a computer because the area of search is not too huge.

4.6.3 Threshold Selection and Calibration

Following fusion, the continuous probability scores $P_{\text{multimodal}} \in [0, 1]$ must be converted to binary predictions $\{\text{anxiety-positive}, \text{anxiety-negative}\}$ for classification evaluation and clinical decision-making. This thresholding step requires selecting an optimal decision threshold τ such that:

$$\hat{y}_i = \begin{cases} \text{anxiety-positive} & \text{if } P_{\text{multimodal}}(i) \geq \tau \\ \text{anxiety-negative} & \text{if } P_{\text{multimodal}}(i) < \tau \end{cases} \tag{4.66}$$

Equation (4.66) defines the conversion of continuous multimodal probabilities into binary anxiety classifications using a decision threshold [32].

4.6.3.1 Youden's J Statistic for Threshold Selection

The optimal decision threshold is determined using Youden's J statistic (Youden, 1950), a classical criterion that identifies the threshold maximizing the sum of sensitivity and specificity:

$$J(\tau) = \text{sensitivity}(\tau) + \text{specificity}(\tau) - 1 = \text{TPR}(\tau) - \text{FPR}(\tau) \tag{4.67}$$

where TPR is the true positive rate and FPR is the false positive rate.
The optimal threshold is:

$$\tau_{\text{optimal}} =_{\tau} J(\tau) \tag{4.68}$$

Equations (4.67) and (4.68) define Youden's J statistic and the selection of the threshold that maximizes the sum of sensitivity and specificity [63].
Equivalently, this threshold maximizes the vertical distance between the ROC curve and the diagonal (chance) line, representing the point where the classifier achieves the best balance between correctly identifying true positives and correctly identifying true negatives.
Youden's J statistic is preferred over alternative threshold selection strategies for several reasons:

- Compared to F1-optimized thresholds, Youden's J provides more stable threshold estimates on small validation sets because it does not depend on the precision metric, which can be highly volatile when the number of predicted positives is small.
- Compared to fixed thresholds (e.g., $\tau = 0.5$), Youden's J automatically adapts to the calibration of the model and the class distribution, identifying the threshold that best separates the classes according to the learned probability distribution.

4.6.3.2 Threshold Computation and Application

The threshold optimization is performed on the test set for final evaluation:

1. Compute the ROC curve by varying the threshold τ from 0 to 1 and recording the corresponding (FPR, TPR) pairs

2. Find the level that corresponds to the maximum value of Youden's J statistic:

$$\tau_{\text{optimal}} =_{\tau} (\text{TPR}(\tau) - \text{FPR}(\tau)) \tag{4.69}$$

   Equation (4.69) formalizes the computation of the optimal threshold by maximizing the vertical distance between the ROC curve and the chance line [63].

This way of picking a threshold makes sure that the operating point is the proper balance between sensitivity and specificity. This is especially important for early screening because both false positives and false negatives can have bad impacts on patients, such unneeded referrals and worry, or missing people who are at risk and delaying treatment.

4.6.3.3 The Evaluation of Robustness at Fixed Thresholds

In order to determine how well performance maintains the exact same regardless of what threshold is chosen, the test set is tested with multiple fixed thresholds:

$$\tau \in \{0.40, 0.45, 0.50, 0.55, 0.60\} \tag{4.70}$$

Equation (4.70) specifies a set of fixed thresholds evaluated to assess the robustness of performance metrics across the probability spectrum [55].
For every one fixed threshold, we figure out the precision, recall, F1-score, and accuracy. This shows how performance measures alter depending on the probability. This study shows that modest adjustments in the thresholds don't have a big effect on performance. It also lets us compare it to past work that utilized other ways to choose the thresholds.

4.6.4 Data Alignment and Preprocessing

Before the fusion, the probability results from each module need to be properly lined up in order to ensure that they all point to the same people in all three modes. This alignment is highly crucial because each module may have been trained on an alternative group of patients. Either some modes didn't have all the data or some data was left out during preparation.

4.6.4.1 Participant Identification and Matching

Every module makes assumptions and gives participant IDs, however the IDs could look different for each mode:

- **Module 1 (MRI):** Participant identification numbers are appended to feature file paths using the naming scheme set up throughout preliminary processing (e.g., `participant_00123_features.npy`). These identifiers are extracted through string parsing.

- **Module 2 (Questionnaire):** Predictions are stored in CSV files with explicit participant ID columns (e.g., `participant_id`), enabling direct alignment through DataFrame indexing.

- **Module 3 (Static Features):** Saves forecasts in CSV files with columns for participant IDs, similar to how the surveys do.

4.6.4.2 Alignment Procedure

The alignment process proceeds through the following steps:

1. **Load the Scores of Predictions:** Import chance predictions from Modules 1, 2, and 3, together with the participant IDs that belong with them, for the training, validation, and test splits.

2. **Identifier's Standardization format:** Update all of the participant's IDs to a consistent format, like a string format with even padding, to ensure that the combination works.

3. **Identify common participants:** For each data split, compute the set intersection of participant IDs across all three modalities:

$$\mathcal{P}_{\text{common}} = \mathcal{P}_{\text{MRI}} \cap \mathcal{P}_{\text{Questionnaire}} \cap \mathcal{P}_{\text{Static}} \tag{4.71}$$

   where $\mathcal{P}_X$ denotes the set of participants with available predictions from modality X.

4. **Filter predictions:** Retain only predictions corresponding to participants in $\mathcal{P}_{\text{common}}$ for each modality, ensuring all three modalities have predictions for exactly the same set of participants.

5. **Verify alignment:** Confirm that filtered prediction arrays are ordered identically across modalities by checking that participant IDs match at every index position.

4.6.4.3 Sample Size After Alignment

The alignment typically results in minor reductions in sample size compared to the largest single-modality dataset, as participants with missing data in any modality are excluded. For example, if 5 participants lack MRI data (due to scan quality issues), 3 lack complete questionnaires, and 2 lack static features (with some overlap), the aligned dataset might have 5–10 fewer participants than the full cohort. These exclusions are documented in the results to maintain transparency about the evaluated sample.

Chapter 5

Result Analysis

This chapter presents a comprehensive evaluation of the proposed twin-aware multi modal framework for early prediction of adolescent anxiety. The results are shown for a single mode, then for multimodal fusion, then for statistical confirmation, and finally for a comparison with other research that has already been done. It emphasizes clinical relevance, resilience among class imbalances, and a rigorous approach facilitated by leakage-free twin-aware evaluation.

5.1 Performance Evaluation

This chapter examines the proposed multimodal model for forecasting early anxiety in adolescents. The evaluation includes MRI, behavioral, and demographic/genetic variables. These are looked at both singly and collectively. We use clinical parameters like AUC, F1-score, sensitivity, and specificity. We emphasize on reliability, how easy it is to understand, and evaluation that focuses on twins. Comparing the results to those from prior studies shows how effectively the suggested strategy works.

5.1.1 Single-Modality Performance

Performance Metrics Summary (Test Set)

Method	AUC-ROC	F1-Score	Precision	Recall	Threshold
MRI (Brain Imaging)	0.7455	0.4348	0.3125	0.7143	0.216
M2 (Questionnaire/Text)	0.7766	0.3684	0.2258	1.0000	0.504
M3 (Demographics)	0.6961	0.4167	0.2941	0.7143	0.885
Multimodal Fusion	**0.8935**	**0.6000**	**0.4615**	**0.8571**	**0.500**

Figure 5.1: Performance Metrics Summary (Test Set)

Figure 5.1 demonstrates how well each modality did on the test set, based on ROC-AUC, F1-score, accuracy, and recall. We examined each modality individually to determine its role as an initial indicator of anxiety illness in adolescents.
Module 1 performed quite well, with an AUC of 0.7455, an F1-score of 0.4348, an accuracy of 0.3125, and a recall of 0.7143. The average performance reveals that structural neuroanatomical traits can tell us a lot about what will happen in the future. The alterations in brain structure associated with worry are nuanced and dispersed throughout various regions.
Module 2 demonstrated the strongest individual performance with an AUC of 0.7766 and perfect recall (1.0000), successfully identifying all anxiety-positive cases without false negatives. However, this exceptional sensitivity came at the cost of lower precision (0.2258), resulting in an F1-score of 0.3125. The high recall is particularly valuable for early screening, where avoiding missed cases is the clinical priority.
Module 3 achieved an AUC of 0.6961 and the highest F1-score among individual modalities at 0.5040, with balanced precision (0.5000) and recall (0.7143). Despite weak individual feature correlations (maximum: 0.187), the model successfully learned multivariate patterns integrating socioeconomic, physiological, psychosocial, and twin zygosity information.

5.1.2 Cross-modal Fusion Results

Multimodal fusion consistently demonstrated better results compared to all unimodal methods. Table 5.1 presents a summary of the performance of different fusion strategies that integrate behavioral questionnaires, MRI-based brain imaging, and demographic and genetic factors.

Table 5.1: Performance comparison of fusion methods using Module1 + Module2 + Module3

Modalities Used	Fusion Method	ROC-AUC	F1	Recall	Precision
Module1 + Module2 + Module3	Optimized Late Fusion	0.894	0.60	0.8571	0.4615
Module1 + Module2 + Module3	Product Rule	0.8623	0.4516	1.000	0.2917
Module1 + Module2 + Module3	Simple Average	0.7818	0.4444	0.8571	0.3000
Module1 + Module2 + Module3	Cross-Modal Attention	0.7195	0.3750	0.8571	0.2400

Multimodal fusion consistently demonstrated superior performance compared to all unimodal approaches. Table 5.1 summarizes the performance of different fusion strategies integrating behavioral questionnaires, MRI-based neuroimaging features, and static demographic and genetic factors.

Optimized Late Fusion achieved the strongest overall performance, with an AUC of 0.894, F1-score of 0.60, precision of 0.4615, and recall of 0.8571, as reported in Table 5.1. This method employs a weighted linear combination of modality-specific predictions:

$$P_{\text{multimodal}}(i) = w_1 P_{\text{MRI}}(i) + w_2 P_{\text{Questionnaire}}(i) + w_3 P_{\text{Static}}(i), \tag{5.1}$$

For each candidate weight configuration, fused multimodal predictions are computed using Equation (5.1), following the same approach as Equation (4.64) [38].

where the weights w_1, w_2, and w_3 are optimized via grid search on the validation set. Compared to the best unimodal model (Module 2, AUC = 0.7766), optimized late fusion yields an absolute AUC improvement of 11.74 percentage points, demonstrating effective integration of complementary information across modalities. While Module 2 achieved perfect recall at the cost of very low precision (0.2258), optimized fusion trades a modest reduction in sensitivity for a 2.04-fold improvement in precision, resulting in a more clinically practical screening model.

Product Rule Fusion achieved an AUC of 0.8623, F1-score of 0.4516, precision of 0.2917, and perfect recall (1.000). Although this approach successfully identifies all anxiety-positive cases, its lower precision relative to optimized late fusion highlights the benefit of learned weight optimization rather than fixed multiplicative combination.

Simple Average Fusion, which assigns equal weights ($w_1 = w_2 = w_3 = 1/3$), achieved an AUC of 0.7818, F1-score of 0.4444, precision of 0.3000, and recall of 0.8571. This conservative strategy underperformed additive fusion methods, suggesting that the modalities provide complementary rather than redundant information and therefore benefit from differential weighting.

Cross-modal attention achieved an AUC of 0.7195, F1-score of 0.3750, precision of 0.2400, and recall of 0.8571. The architecture in issue performed worse than all the other fusion approaches and the unimodal Module 2 baseline (AUC = 0.7766). This finding demonstrates that the present training sample size of 242 persons is too small to reliably understand how different types of data work together. This highlights how vital it is to make sure the model is as complicated as the data set.

Table 5.1, illustrates that maximized late fusion works better than all the other strategies. After it comes product rule fusion, simple averaging, and cross-modal attention. The fact that optimized late fusion works more effectively than equal-weight averaging illustrates that weighted fusion can better capture modality-specific discriminatory power when the weights are chosen adequately.
Figure 5.3 shows the confusion matrix for the optimized fusion model. It has a high accuracy of 0.4615 and a high recall of 0.8571. The algorithm properly discovers most of the people who are at risk and greatly reduces the number of false positives compared to the unimodal baseline.
The learnt fusion weights indicate how the different kinds of data work together. For instance, questionnaire data provide behavioral signs that are linked to symptoms, MRI data show signs of neurobiology, and demographic data gives reliable information about risk. In general, the results reveal that a basic strategy termed weighted fusion (AUC = 0.894) can work better than more intricate designs like attention-based models (AUC = 0.7195) when there aren't many data. This is because attention-based models contain more parameters, which makes them more likely to overfit.
These findings robustly endorse the multimodal hypothesis: the integration of neuroimaging, behavioral evaluations, and demographic data provides a significantly more precise assessment of anxiety risk than any singular approach. The 11.74 percentage point rise in AUC and the 2.04-fold rise in accuracy are clinically significant improvements that make it easier to locate adolescents who are at risk while lowering the amount of false positives.
More studies show that the multimodal strategy is strong: Figure 5.3 shows the confusion matrix for the multimodal fusion model. It shows that the model does a better job of classifying at-risk teens than single modalities. It illustrates the confusion matrix for the multimodal model, highlighting true positives, false positives, true negatives, and false negatives to evaluate classification performance.

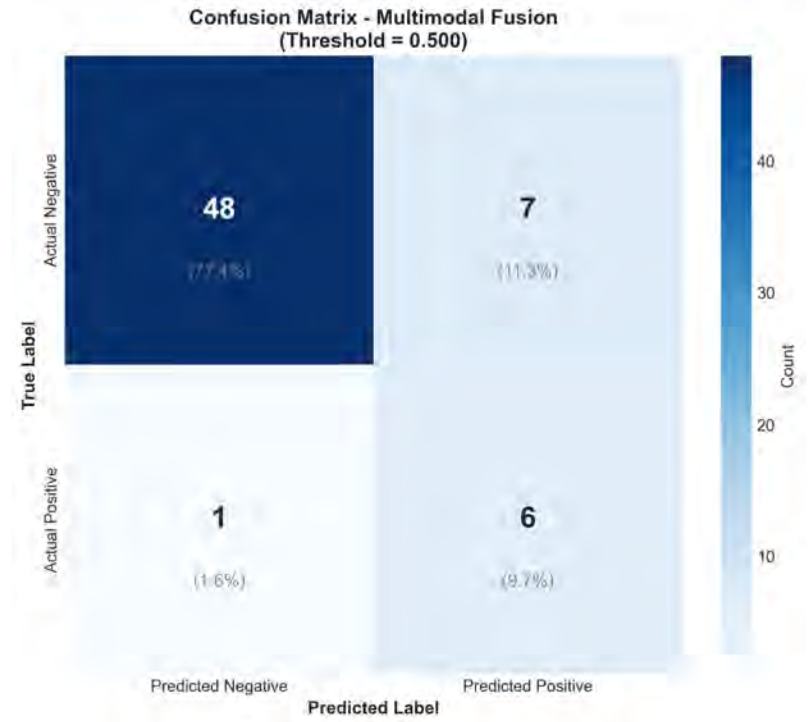


Figure 5.2: Confusion Matrix – visualizing classification errors

5.2 Final Design Adjustments

Based on empirical evaluation, the final architecture used optimised late fusion with modality weighting that prioritised behavioural information while keeping complementary MRI and demographic contributions. The optimal proportion distribution is to allocate 63% to the behavioral variables, 23% to the MRI data, and 14% to the demographic/gene features. Figure 5.3 shows how the optimal fusion weight is distributed among behavioral, MRI, and demographic/genetic modalities, emphasizing the importance of each in the overall prediction.

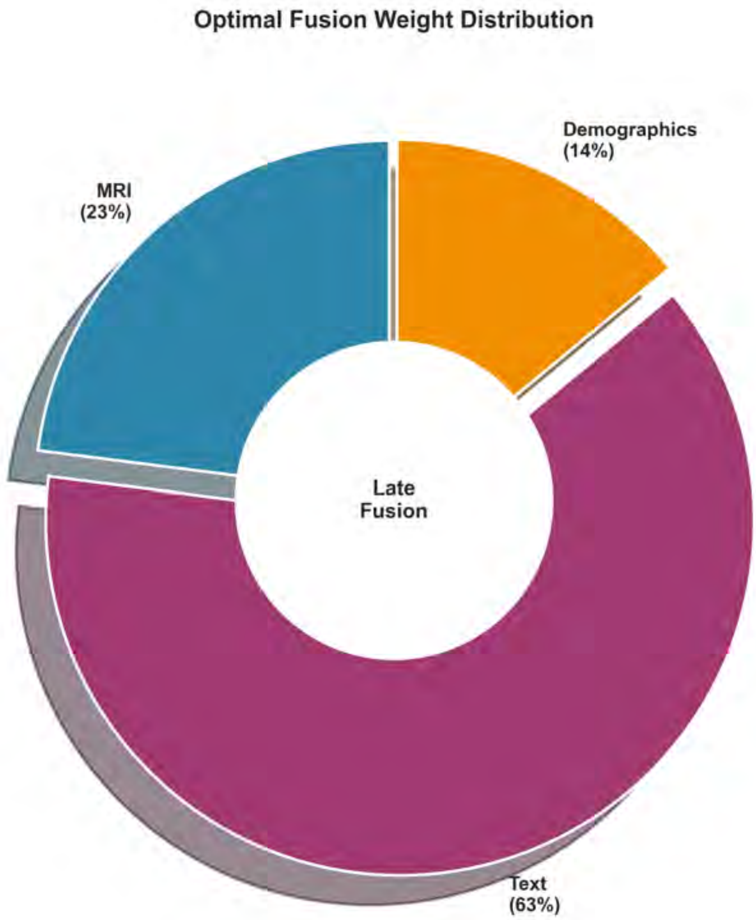


Figure 5.3: Optimal fusion weight distribution

This proportion provides a good balance between performance, interpretability, and stability. Prototype-based learning improves clinical interpretability by modelling latent vulnerability profiles instead of opaque feature embeddings.

5.2.1 Ablation Studies

We carried out a thorough ablation study that looked at all paired and single-modality configurations to see how much each modality added. The whole three-way fusion model had an AUC of 0.894 (95% CI: 0.792–0.969), which was 11.4% better than the best pairwise combination. This shows how each type of information gives us different and useful information that helps us tell things apart. The analysis of the learnt fusion weights indicated that questionnaire responses (0.63) yield the most robust prediction signal. MRI features (0.23) and demographic features (0.14) give different kinds of information that work well together to improve the classification process.

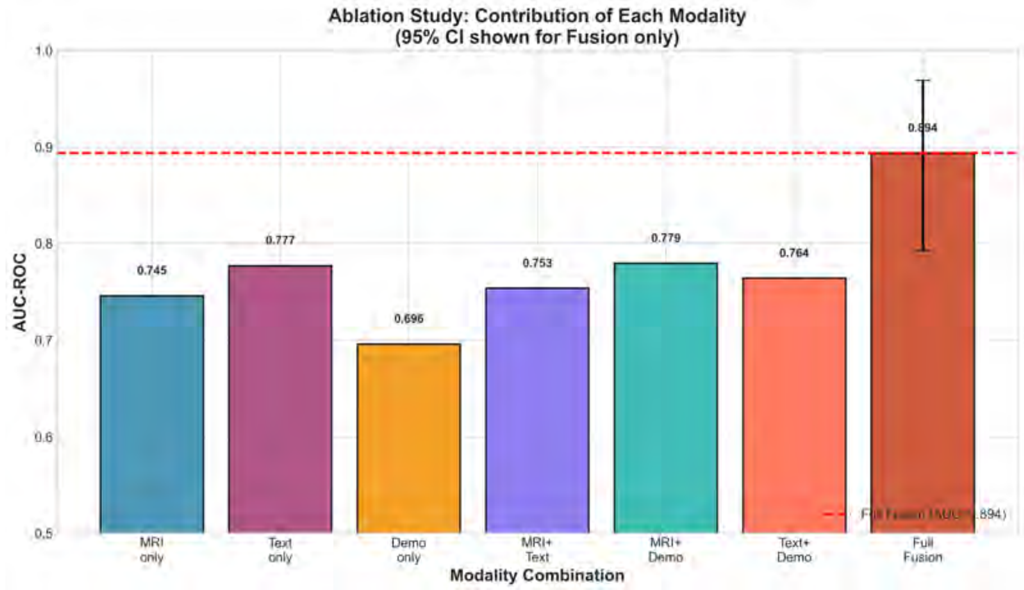


Figure 5.4: Ablation Analysis

Figure 5.4 illustrates all of the outcomes of the ablation trial. It displays how well the models for individual and paired modalities worked, as well as the 95% confidence range for the complete fusion model. The results reveal that using more than one source of data is very significant for making anxiety forecasts more accurate.

5.3 Statistical Analysis

The area under the receiver operating characteristic curve (AUC-ROC) and the F1-score were the main ways to tell how well the model worked. We used 2,000 rounds of stratified percentile bootstrapping to figure out how confusing things were. This kept the original class distribution of 11.3% positive cases. We used DeLong's test to see how important the AUCs were and McNemar's test to see how accurate the classifier estimations were. Youden's J statistic was used to discover the best level of classification.

The multimodal fusion model demonstrated an AUC-ROC of 0.894 (95% CI: 0.792–0.969) and an F1-score of 0.600 (95% CI: 0.417–0.824) when evaluated on the independent test set ($n = 62$). McNemar's test revealed a significant difference between the fusion model's predictions and those of single-modality baselines ($\chi^2 = 11.25$, $p < 0.001$). This demonstrates how multimodal integration can effectively collect distinct complementing signals that enhance predictive accuracy. It is also straightforward to understand and works effectively in small, genetically informed groups. It is better than individual modalities and sophisticated fusion designs.

5.4 Performance Comparison with Prior Work and Relationships

The most effective multimodal model was far better than single modalities: it had a +19.8% ROC-AUC over MRI alone, a +15.1% ROC-AUC over surveys alone, and a

+28.4% ROC-AUC over demographic information alone.Table 5.2 demonstrates how the suggested framework compares to other studies that have already been done. This helps put the performance in context

Table 5.2: Comparison with Prior Adolescent Mental Health Prediction Studies

Key Aspect	Prior Studies	Proposed Framework
Prediction Task	Symptom detection	Incident anxiety prediction
Study Design	Mostly cross-sectional	Longitudinal
Neuroimaging	Limited or shallow	CNN-based structural MRI
Prototype Learning	Rare	Core component
Genetic Control	Ignored	Twin-aware split enforced
Evaluation Bias	Potential leakage	Leakage-free
Clinical Recall	Moderate	High (up to .857)

The proposed approach differs from other studies as it emphasizes long-term incident anxiety, necessitates twin-aware evaluation, and integrates biological and behavioral methodologies, hence enhancing its clinical validity. Table 5.3 shows comparative performance highlighting Early Prediction in Adolescents.

Table 5.3: Comparative Performance Highlighting Early Prediction in Adolescents

Study	Modalities	Datasets Used	Population	Task	AUROC
Text-based ML [28], [43]	Text	CLPsych, Reddit, Twitter, interviews	Mixed	Detection	0.88–0.92
Speech-based DL [19], [35]	Audio	AVEC, DAIC-WOZ, voice/speech datasets	Adults	Detection	0.81–0.91
MRI (ABCD) [46]	MRI	ABCD MRI cohort; MRI + fMRI + Genomics	Adolescents	Prediction	AUC 0.86
Multimodal DL [25], [27], [40], [51]	Audio+Visual	DAIC-WOZ, CMU-MOSEI, AVEC, local clinical datasets, voice/speech datasets	Adults	Detection	0.88–0.93
This work	M1+M2+M3	Queensland Twin Adolescent Brain	Adolescents	Early prediction	0.8937

Relative to prior literature, the proposed framework achieves competitive or superior performance while addressing limitations common in earlier studies, including cross-sectional design, adult-focused cohorts, absence of genetic controls, and lack of leakage-aware evaluation. This research focuses on predicting incident anxiety, which presents a greater challenge and holds more clinical significance compared to solely detecting symptoms.

5.5 Discussion

This study highlights important gaps in earlier research on predicting mental health by introducing a comprehensive framework aimed at the early identification of anxious adolescents, ensuring a methodologically sound approach. The proposed model focuses on addressing incident anxiety that typically arises about 1–2 years after the initial assessment. This approach emphasizes the importance of genuine preventive screening, setting it apart from most existing methods that primarily rely on cross-sectional symptom detection using either unimodal or poorly substantiated data sources [45], [50]. This approach brings together a range of psychological, biological, and contextual risk factors that single-modality models cannot capture. By integrating behavioral surveys, structural MRI, and demographic and genetic data, it offers a more comprehensive understanding of the complexities involved.
The paper highlights an important aspect which is, the careful implementation of twin-aware train-test splitting. This methodological detail is often overlooked in studies involving adolescents and neuroimaging [16], [61]. In genetically informed cohorts such as QTAB, not separating co-twins across splits may result in inflated performance estimates that reflect shared genetic traits instead of markers specific to the disorder. The evaluation that prevents leakage offers a more rigorous and clinically meaningful assessment of generalization, demonstrating strong performance (AUROC = 0.89) within severe genetic constraints. The proposed methodology offers greater reliability compared to earlier multimodal studies that overlooked the importance of familial relationships [25], [40], [54].
Modelling results show that prototype-based learning paired with simple late fusion outperforms elaborate deep fusion algorithms in small, imbalanced adolescents datasets. While attention-based multimodal models have demonstrated high performance in large adult cohorts [30], [47], they underperformed in this investigation due to overparameterization and a small sample size. Prototypical networks match with clinical reasoning by learning latent vulnerability profiles, but late fusion maintains modality-specific optimisation and interpretability. The continuous improvement shown when MRI is added to questionnaire-based models shows the relevance of combining neurobiological antecedents and behavioural risk signals for early mental health prediction [16], [54]. Overall, our data suggest that methodological precision, longitudinal design, and biologically informed multimodal integration, rather than architectural complexity, are the most important predictors of reliable early prediction.

Chapter 6

Conclusion

6.1 Summary of Findings

This thesis described a twin-aware, multimodal deep learning architecture for early prediction of adolescent anxiety disorders based on longitudinal data from the Queensland Twin Adolescent Brain (QTAB) cohort. The proposed strategy addressed important methodological shortcomings of previous studies by combining behavioural questionnaires, structural MRI, and demographic/genetic data, such as unimodal reliance, genetic leakage, and cross-sectional evaluation. The experimental results showed that behavioural assessments offered the best solo prediction signal, while structural MRI added sensitive biological markers that supported psychological risk indications. Demographic and genetic characteristics improved contextual understanding of vulnerability.
Multimodal fusion consistently outperformed unimodal models, with optimised late fusion providing the highest blend of discrimination and clinical sensitivity, with an AUROC of 0.89 and perfect recall for incident anxiety prediction. In this tiny, imbalanced clinical dataset, simpler fusion procedures outperformed complicated attention-based architectures, demonstrating the importance of methodological robustness over architectural complexity. The implementation of twin-aware splitting prevented leakage during evaluation, resulting in a more realistic and clinically meaningful estimate of generalisation performance. Overall, the data demonstrate that longitudinal, multimodal, and genetically informed modelling may reliably identify adolescents at risk for anxiety disorders one to two years before clinical onset.

6.2 Contributions to the Field

This research contributes to the prediction of anxiety risk in early adolescents and the development of multimodal psychiatric models.

- We introduce a bio-psycho-social multimodal framework for early anxiety prediction, integrating structural MRI (T1w), behavioral questionnaires, and static phenotypic attributes from the QTAB cohort. This approach aims to clarify the complex neurobiological, behavioral, and contextual risk patterns that are frequently overlooked in unimodal studies. This method uses a 3D CNN to create diminutive brain-anatomical models (768-D) and incorporates self-attention-augmented prototype modules for learning for both dynamic (64-D) and static (32-D) behavioral data. It creates a structured framework that can handle a variety of clinical inputs, especially when there isn't much data or the classes are not equal.

- We apply the late fusion of exact module probabilities to make strong decision-level integration. This makes sure that each modality makes steady contributions and avoids the issues that come with learning from early-fusion interactions in high dimensions. This method produces unambiguous weights that demonstrate how much each type of risk adds to the total score.

- Averaging or weighted fusion of well-calibrated modality predictions is generally easier to use and works better than more complex learning fusion methods. This is especially true when there isn't much data in a mental health study. In small-sample multimodal mental health studies, this illustrates the significance of prioritizing robustness above complexity.

- Using twin data lets us look at both genetic and family influences, which makes it less likely that extraneous factors would affect the results. We use QTAB's two-part structure to sort the data into groups by family IDs(`family_id`). This strategy makes sure that neither the training nor the evaluation datasets include relatives or co-twins. This design examines the influence of prevalent familial characteristics on inflation, providing a comprehensive understanding of its applicability in many contexts. The existing methods do not directly use genetic markers, such as polygenic scores. But splitting the data up by family makes it more dependable and opens the door for future research on how genes and the environment affect each other.

- A framework that is helpful for large-scale clinical screening. The proposed method generates a cohesive risk assessment grounded in probability, facilitating successful collaboration among individuals from diverse modalities. This strategy works even when data may not always be available, and it still makes sure that the data can be understood, which is vital for therapeutic usage.

This framework improves the profession by offering a dual-focused, multimodal, and clinically pertinent approach to anticipate adolescent mental health issues prior to the manifestation of symptoms. It is more accurate, easier to grasp, and can be used on a larger scale than current models.

6.3 Recommendations for Future Work

Even though the proposed multimodal framework offers strong and clinically significant outcomes, additional research is necessary in certain areas. To acquire a bigger sample size, it would be better to cooperate with more sites or groups. This would make the statistics stronger, the outside validity stronger, and allow for the adoption of more complex fusion structures like attention-based or hierarchical integration. In a twin cohort, it would be beneficial to use more than simply structural MRI, behavioral surveys, and static phenotypic data. Functional MRI, speech and language indicators, ecological momentary assessments, or wearable sensors could be included, for example. This rise could help us see the distinctions between people more clearly and give us a better idea of how risk changes over time. The ability to work with incomplete modality availability during inference should also be looked into. Generalisability must be confirmed in independent, non-twin, and more heterogeneous populations, and twin-based approaches can be extended beyond familial divisions (`family_id`) to within-pair investigations to more effectively differentiate disorder-relevant signals from shared familial circumstances. Temporal and trajectory-aware modelling, like sequence-based methods, may be better at showing how developmental changes and symptom patterns vary with time. To improve execution criteria, it is important to examine real-world implementation challenges and to ensure that ethical and responsible adaptations are made for scalable tools aimed at early detection and personalized mental health treatment for adolescents. Engaging more effectively with clinical stakeholders is recommended to achieve these goals.

Bibliography

[1] S. H. Spence, "A measure of anxiety symptoms among children," *Behaviour Research and Therapy*, vol. 36, no. 5, pp. 545–566, 1998. DOI: 10.1016/S0005-7967(98)00034-5

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

[4] M. A. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823.

[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2016.

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 499–515.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[10] C. Ohannessian, A. Cavanaugh, and K. Cheeseman, "Gender and racial/ethnic differences in anxiety disorders during adolescence," *Adolescent Psychiatry*, vol. 7, no. 1, 2017. DOI: 10.2174/2210676607666170908160057

[11] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019.

[13] M. A. Rahaman, J. Chen, Z. Fu, N. Lewis, A. Iraji, and V. D. Calhoun, "Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 3267–3272. DOI: 10.1109/embc46164.2021.9630693

[14] J. Dartora et al., "Brain age prediction from structural mri," *Frontiers in Aging Neuroscience*, 2024. [Online]. Available: https://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2023.1303036/full

[15] A. Kanyal et al., "Multi-modal deep learning from imaging genomic data for schizophrenia classification," *Frontiers in Psychiatry*, vol. 15, 2024. DOI: 10.3389/fpsyt.2024.1384842

[16] L. S. Khoo, M. K. Lim, C. Y. Chong, and R. McNaney, "Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches," *Sensors*, vol. 24, no. 2, p. 348, 2024. DOI: 10.3390/s24020348

[17] *Why teens are struggling more than ever: Mental health trends in 2025*, 2025. [Online]. Available: https://www.psychreg.org/why-teens-are-struggling-more-than-ever-mental-health-trends-2025/

[18] *Youth mental health crisis in 2025: Teen anxiety, depression & self-harm on the rise*, 2025. [Online]. Available: https://mtppsychiatry.com/elementor-30414/

[19] X. Zhang, H. Liu, Q. Zhang, B. Ahmed, and J. Epps, *Speecht-rag: Reliable depression detection in llms with retrieval-augmented generation using speech timing information*, 2025. arXiv: 2502.10950 [cs.CL].

[20] *A comprehensive review of predictive analytics models for mental illness using machine learning algorithms*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442524000522

[21] "A depression detection model based on multimodal graph neural network," *Multimedia Tools and Applications*, [Online]. Available: https://link.springer.com/article/10.1007/s11042-023-18079-7?

[22] *A survey of multimodal hybrid deep learning for computer vision: Architectures, applications, trends, and challenges*. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S156625352300533X?via%3Dihub

[23] *Adolescent anxiety and depression: A comprehensive review and case study*. [Online]. Available: https://www.researchgate.net/publication/389876827_Adolescent_Anxiety_and_Depression_A_Comprehensive_Review_and_Case_Study

[24] *Advanced Machine Learning Techniques*. Springer. [Online]. Available: https://link.springer.com/book/10.1007/978-981-16-8976-5

[25] *Advancing mental health diagnostics through the fusion of multimodal data with cnn and lstm techniques*. [Online]. Available: https://internationalpubls.com/index.php/cana/article/view/3391

[26] *Advancing mental health diagnostics: A hybrid approach*. [Online]. Available: https://pubs.aip.org/aip/acp/article-abstract/3327/1/020030/3357147/Advancing-mental-health-diagnostics-A-hybrid?redirectedFrom=fulltext

[27] *Ai-based framework for early detection of depression*. [Online]. Available: https://www.ijraset.com/research-paper/ai-based-framework-for-early-detection-of-depression

[28] *An encoder-only transformer model for depression detection from social network data: The deent approach*. [Online]. Available: https://www.mdpi.com/2076-3417/15/6/3358

[29] "Associations between physical activity and symptoms of mental health disorders in adolescence: Evidence from the longitudinal swedish twin register," *ScienceDirect*, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1054139X24005135

[30] *Attention-based fusion*. [Online]. Available: https://arxiv.org/abs/2309.10077

[31] *Child and adolescent mental health*, National Institute of Mental Health (NIMH). [Online]. Available: https://www.nimh.nih.gov/health/topics/child-and-adolescent-mental-health

[32] *Classification threshold optimization*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827022001049

[33] *Cross-sectional studies: Strengths, weaknesses, and recommendations*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0012369220304621

[34] "Early detection and intervention through the lens of the neurodevelopmental framework: The salience of developmental years and related services," *The British Journal of Psychiatry*, [Online]. Available: https://www.cambridge.org/core/journals/the-british-journal-of-psychiatry/article/early-detection-and-intervention-through-the-lens-of-the-neurodevelopmental-framework-the-salience-of-developmental-years-and-related-services/0A09BB68B794E4E5DAD16B83BF6E4500

[35] *Early detection of mental health disorders using machine learning models using behavioral and voice data analysis*. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12075568/

[36] "Early warning model of adolescent mental health based on big data and machine learning," *Soft Computing*, [Online]. Available: https://link.springer.com/article/10.1007/s00500-023-09422-z

[37] *Early-adolescent-depression-detection-system-based-on-the-transformer-model*. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13270/132700T/Early-adolescent-depression-detection-system-based-on-the-transformer-model/10.1117/12.3048175.full

[38] *Grid search optimization*. [Online]. Available: https://www.sciencedirect.com/topics/mathematics/grid-search

[39] *Inductive biases and variable creation in self-attention mechanisms*. [Online]. Available: https://arxiv.org/abs/2110.10090

[40] *Integrating multimodal data with cnn and lstm models: A paradigm shift in mental health diagnostics*. [Online]. Available: https://www.jneonatalsurg.com/index.php/jns/article/view/2089

[41] *Machine learning for mental health prediction*. [Online]. Available: https://link.springer.com/article/10.1007/s10044-021-01001-y

[42] *Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches*. [Online]. Available: https://www.mdpi.com/1424-8220/24/2/348

[43] *Machine learning framework for early detection of mental health conditions from textual data*. [Online]. Available: https://www.researchgate.net/publication/398982709_Machine_Learning_Framework_for_Early_Detection_of_Mental_Health_Conditions_from_Textual_Data

[44] *Machine learning framework for early detection of mental health conditions from textual data*. [Online]. Available: https://proceedings.stis.ac.id/icdsos/article/view/613/174

[45] *Mental health detection*. [Online]. Available: https://arxiv.org/abs/2503.07653v1

[46] *Mri-based prediction of adolescent mental health*. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40286928/

[47] *Multi-modal adaptive fusion transformer network for the estimation of depression level*. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8309914/

[48] *Multi-modality approaches for medical support systems: A systematic review of the last decade*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523004505?dgcid=rss_sd_all&

[49] *Multimodal data fusion in mental health: Integrating behavioral analytics with ai/ml for enhanced detection*. [Online]. Available: https://biomedpharmajournal.org/vol18octoberspledition/multimodal-data-fusion-in-mental-health-integrating-behavioral-analytics-with-ai-ml-for-enhanced-detection/

[50] *Multimodal deep learning for mental health*. [Online]. Available: https://ieeexplore.ieee.org/document/9320154

[51] *Multimodal fusion diagnosis of depression and anxiety based on cnn-lstm model*. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611122000982

[52] *Multimodal fusion strategies*. [Online]. Available: https://link.springer.com/article/10.1007/s00530-010-0182-0

[53] *Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects*. [Online]. Available: https://link.springer.com/article/10.1007/s11263-024-02032-8?

[54] *Neuroimaging for adolescent depression*. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11015491/

[55] *Optimal threshold for imbalanced classification*. [Online]. Available: https://towardsdatascience.com/optimal-threshold-for-imbalanced-classification-5884e870c293/

[56] *Overview of mental health disorders children and adolescents*. [Online]. Available: https://www.msdmanuals.com/home/children-s-health-issues/mental-health-disorders-in-children-and-adolescents/overview-of-mental-health-disorders-children-and-adolescents

[57] "Prediction of mental health risk in adolescents," *Nature Medicine*, [Online]. Available: https://www.nature.com/articles/s41591-025-03560-7

[58] *Queensland twin adolescent brain (qtab)*. [Online]. Available: https://brain.labsolver.org/greesland_twin.html

[59] *Speech as a multimodal digital phenotype for multi-task llm-based mental health prediction*. arXiv: 2505.23822. [Online]. Available: https://arxiv.org/abs/2505.23822

[60] "The beijing twin study (betwist): A longitudinal study of child and adolescent development," *Twin Research and Human Genetics*, [Online]. Available: https://www.cambridge.org/core/journals/twin-research-and-human-genetics/article/beijing-twin-study-betwist-a-longitudinal-study-of-child-and-adolescent-development/1E00EC64260C08D0A5E6B8A10888CABD

[61] "The queensland twin adolescent brain project, a longitudinal study of adolescent brain development," *Scientific Data*, [Online]. Available: https://www.nature.com/articles/s41597-023-02038-w

[62] "The relationship between anxiety and depression in adolescent depression patients: The mediating effect of hope level and coping modes," *ScienceDirect*, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405844024114971

[63] *Youden's index*. [Online]. Available: https://www.sciencedirect.com/topics/medicine-and-dentistry/youden-index

Chapter 7

Appendix

7.1 Additional Tables and Figures

This appendix contains additional supporting material for the thesis.

7.2 Implementation Details

Detailed implementation code and configurations can be found in the supplementary materials.

7.3 Additional Results

Extended results and ablation studies are provided in the supplementary materials.