Olesia Prykhodko

”MACHINE LEARNING MODELS FOR PREDICTING ATTENTION DEFICIT HYPERACTIVITY DISORDER SYMPTOMS IN YOUTH”

BBT.BI.300 Project Work in Biomedical Informatics

Project work report

|  |  |  |
| --- | --- | --- |
| Medicine and Health Technology  Project supervisor: Milla Juutenin  April, 2025 |  |  |

ABSTRACT

: Olesia Prkhodko

Project work report

Tampere University

Master's degree in Biomedical Technology, Biomedical Informatics

April, 2025

ADHD is one of the most frequently diagnosed neurodevelopmental disorders in children, primarily characterized by inattention, hyperactivity, and impulsivity. Recent research indicates that ADHD also persists in adulthood; however, it is often underdiagnosed in this population.

This study aims to analyze data from a survey conducted by Singh (2024), which assessed ADHD symptoms among university students in India using a self-report scale. The symptoms in this scale are categorized into three domains: inattention, hyperactivity, and impulsivity.

The primary goal of this research is to develop predictive models based on the collected data. The aim is to classify individuals into one of the three ADHD symptom categories (inattention, hyperactivity, or impulsivity) using machine learning models trained on self-report responses and demographic data. The broader objective of this study is to enhance the accuracy of self-assessment tools for ADHD, addressing the issue of underdiagnosis.

Keywords: ADHD, machine learning, predictive models, self-reported scales.

# USE OF AI IN THESIS

I have utilised AI tools in my thesis:

No

Yes

The AI tools utilised in my thesis and their purposes are described below:

Names and versions of AI tools:

Deep Seek (v1.5)

Quillbot (as of April 2025)

Grammarly (v1.233.0)

Purpose of using AI tools:

AI tools were used to support minor aspects of the report writing process. Grammarly was employed for checking and refining grammar, spelling, and clarity in written sections, ensuring academic tone and language consistency. Quillbot was used for citation generation. DeepSeek was used to brainstorm and explore possible approaches for the data preprocessing pipeline, helping to generate initial ideas, which were then developed, implemented, and critically evaluated by me. It was also used to help structure the *Discussion* and *Limitations* sections in a more coherent and organized manner.

Sections where AI tools were used:

* Language correction and refinement: *Introduction*, *Materials and Methods*, *Discussion & Conclusions*, *Work Evaluation*
* Brainstorming ideas for data preprocessing: *Section 2.2 – Data preprocessing*
* Discussion organization

I acknowledge that I am fully responsible for the entire content of my thesis, including the parts generated by AI, and accept accountability for any violations of ethical standards in publications.

Contents

[1. Introduction 1](#_Toc535516164)

[2. MATERIALS AND METHODS 3](#_Toc535516165)

[2.1 Dataset description 3](#_Toc535516166)

[2.2 Data preprocessing 4](#_Toc535516167)

[3. RESULTS 9](#_Toc535516170)

[4. DISCUSSION & CONCLUSIONS 11](#_Toc535516173)

[5. WORK EVALUATION 13](#_Toc535516173)

[REFERENCES 14](#_Toc535516174)

[Annex 16](#_Toc535516175)

# Introduction

Attention Deficit Hyperactivity Disorder (ADHD) is primarily a neurodevelopmental disorder diagnosed in early childhood, marked by ongoing patterns of inattention and/ or impulsivity. Individuals with ADHD often struggle to maintain everyday activities due to challenges in completing tasks and staying organized. Common accompanying symptoms include excessive activity and restlessness, difficulties initiating quiet tasks, and self-control issues, which manifest as hyperactivity, impulsivity, and inattentiveness. (*ADHD in Adults: 4 Things to know*, 1.04.2025)

ADHD can be diagnosed in early childhood with the help of a healthcare provider. Nevertheless, some individuals may remain undiagnosed for various reasons. First, the milder form of ADHD might not cause additional problems until the person encounters more demanding and attention-required experiences such as work or studies. Secondly, the child’s environment may not have been patient enough to recognize the mental condition. (*ADHD in Adults: 4 Things to know*, 1.04.2025) Sex differences also impact the results of medical examinations. According to the Centers for Disease Control and Prevention, physicians are three times more likely to diagnose boys with the condition due to more visible symptom displays (Rucklidge, 2010).

Symptoms of ADHD may persist into adulthood; however, their manifestation might be somewhat changed. For this reason, the diagnosis relies heavily on behavioral history, which can be gathered from family members, and friends. Information can also be accessed via records from childhood, such as school documents. Additionally, the individual should participate in a clinical interview and undergo a variable psychological test to identify any presence of the condition. Specialist’s assessment is especially needed to exclude other ADHD-mimicking conditions (e.g. oppositional defiant disorder, thyroid dysfunction, sleep impairments, and disorders) (Sadek, 2023) and to determine whether ADHD has symptoms of any existing comorbid disorder (e.g. anxiety, depression, tic disorder ) (Gnanavel *et al.*, 2019).

Recent advances in machine learning (ML) have demonstrated the potential to enhance ADHD screening by identifying complex behavioral patterns that traditional methods may miss (Faraone et al., 2021). While self-report scales like the ASRS-V1.1 provide accessible screening, some studies present evidence of poor validity and reliability of self-screening scales due to the high rate of false positives (McCann & Roy-Byrne, 2004), highlighting the need for more robust approaches. This study investigates whether incorporating demographic variables (e.g., age, gender, education level) alongside ASRS-V1.1 responses can improve the prediction of specific ADHD symptom clusters - a critical consideration given known diagnostic disparities in clinical populations (Rucklidge, 2010).

The multi-label classification was employed to simultaneously predict inattentive, hyperactive, and impulsive symptom domains in the young adult population, reflecting their frequent co-occurrence in clinical presentations (Magnus et al., 2023). The approach also evaluates: (1) the relative contribution of demographic features to symptom prediction accuracy, and (2) the efficacy of synthetic data augmentation (MLSMOTE) in addressing class imbalance (Tarekegn et al., 2021). By benchmarking models including Classifier Chains and Binary Relevance, I aim to test whether reported self-screening scale results and contextual factors can help detect the condition in the young adult patient group while emphasizing these tools should augment, not replace, clinical judgment.

# Materials and methods

## Dataset description

## The dataset was collected from Zenodo and utilized in a research study by Singh et al. (2024), accessed on February 2, 2025. The dataset comprises information from 360 participants, all students of higher educational institutions in Punjab, India. Among them, 228 were female (63.3%) and 132 were male (36.7%), ranging from 17 to 23 year\*\*s\*\* (mean value = 18.83, median 19, standard deviation - 1.22).

## The study included students who had normal or corrected-to-normal vision with an acuity of 20/20 for distance and N6 for near vision, achieved using glasses or contact lenses. Individuals were excluded if they had a history of eye injuries, surgeries, infections, or any diagnosed neurological or psychological conditions such as ADHD, learning disabilities, or mental disorders. Additionally, students undergoing any form of psychological, behavioral, occupational, or vision therapy, as well as those with a history of head trauma or significant general health issues, were not considered for participation.

## Each participant's data includes demographic information (age, gender), family background, parental education, and living arrangements. The distribution of participants' characteristics is presented in graphs (1A-F) below:

|  |  |
| --- | --- |
|  |  |
| Figure 1A. Family type distribution. 1 -Nuclear Family, 2 - Joint/Extended Family, 3 - Single-Parent Family, 4 -Grandparent Family, 5 - Stepfamily | Figure 1B. Living arrangements. 1 - Hosteler, 2 - Day Scholars |
|  |  |
| Figure 1C. Father education level. 1 - Uneducated, 2 - 10th grade, 3 - 12th grade, 4 - Graduation, 5 - Master’s | Figure 1D. Mother education level. 1 - Uneducated, 2 - 10th grade, 3 - 12th grade, 4 - Graduation, 5 - Master’s |
|  | |
| Figure 1F. Sibling’s number distribution | |

## Data preprocessing

The Adult ADHD Self-Report Scale (ASRS) was used for symptom assessment, following the scoring and classification guidelines provided in the ASRS interpretation guide (Buchanan, 2025).

1. **Scores Calculation**

New columns were created to reflect the summed-up scores based on different sections of the ASRS. For part A (questions 1–6) **s**core ranged from 0 to 24, part B (questions 7-18) - from 0 to 48, and a total score ranges from 0 to 72. The figure presented (Figure 2) shows the distribution of calculated total scores, which shows almost a Gaussian distribution.

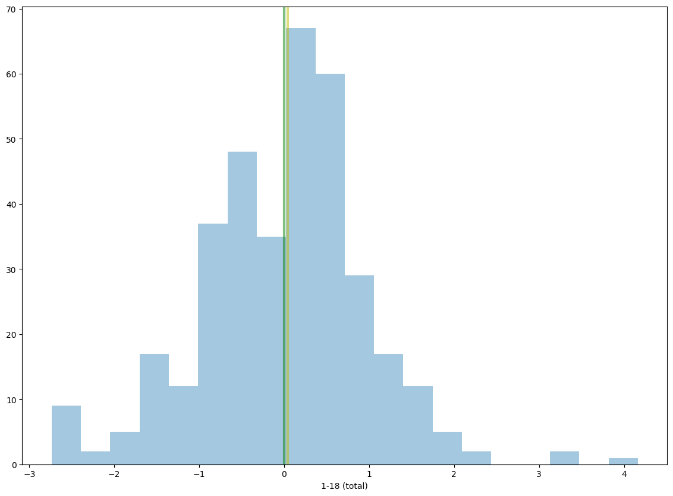


Figure 2. Total score distribution. Mean - green line, median - yellow line.

1. **Subscale Scoring**

Following the original ASRS dichotomous scoring method, responses were recoded into binary values (0 or 1) for three subscales: inattentiveness (Questions 1, 2, 3, 4, 7, 8, 9, 10, 11; Range: 0–9), Hyperactivity (Verbal) (Questions 5, 6, 12, 13, 14; Range: 0–5), Hyperactivity (Motor) (Questions 15–18; Range: 0–4).

To align with the **DSM-5** diagnostic criteria for ADHD subtypes (Substance Abuse and Mental Health Services Administration (US), 15/04/2025), the percentage of endorsed symptoms was analyzed. The following thresholds were applied: inattentive type: inattentiveness subscale ≥6; hyperactive-impulsive type: combined motor and verbal hyperactivity ≥6; combined type: both subscales ≥6**.**

However, due to the small dataset size, this approach resulted in an extremely limited number of ADHD-subtype labels (Hyperactive-Impulsive: **1**, Inattentive: **4**, Combined: **2**), making further subtype analysis unfeasible.

Given the small dataset size and the limited number of ADHD-subtype labels, an alternative method was used to determine the dominant ADHD trait based on symptom prevalence.

Instead of applying strict DSM-5 thresholds, the percentage of endorsed items was calculated for each symptom category (Inattentiveness, Hyperactivity-Verbal, and Hyperactivity-Motor). The symptom with the highest percentage was identified as the dominant trait and assigned a label of 1, while the other two were assigned 0. If two or all three symptoms had the same percentage of endorsed items, they were all assigned 1 instead of selecting a single dominant trait. This approach allowed for a more flexible classification, capturing relative symptom prominence.

1. **Imbalance dataset handling**

The labels were designed with the understanding that ADHD symptoms often overlap in real-life scenarios, meaning a single individual may exhibit multiple symptom types simultaneously (Magnus, Anilkumar, and Shaban, 2023). To address this complexity, the multi-label classification approach was adopted, similar to methods used in classifying psychotic diseases, where patients frequently present with multiple overlapping symptoms (Folorunso et al., 2020).

To assess the distribution of symptom categories within the dataset prior to applying machine learning models, I calculated the prevalence of each label. Specifically, inattentiveness was present in 26.7% of the cases, motor hyperactivity in 18.6%, and verbal hyperactivity in 11.4%. These proportions were computed as the ratio of positive labels to the total number of observations.

These results are providing insight on the noticeable class imbalance, with inattentiveness being the most prevalent symptom. This imbalance is a common challenge in multi-label classification, as it can manifest in various forms, such as imbalance between labels or within label sets. Such imbalances can negatively impact model learning, leading to biased predictions and poor generalization (Tarekegn, Giacobini, and Michalak, 2021). Additionally, the relatively small dataset size (360 instances) increases the risk of overfitting and limits the model's ability to generalize to new data.

To address these challenges, two distinct approaches were evaluated:

1. K-Fold Cross-Validation on the Original Dataset: This approach assesses model performance on the imbalanced dataset without any modifications, providing a baseline for comparison.

2. K-Fold Cross-Validation with Synthetic Data Augmentation: MLSMOTE (Multi-Label Synthetic Minority Oversampling Technique) was used to balance the dataset by generating synthetic samples for underrepresented labels, thereby mitigating the effects of class imbalance. To compare the effectiveness of these approaches, three benchmarking models were employed, in particular, used for the biomedical data (Diakou et al., 2024). As for the comparison metrics precision, recall, F1-score with micro, macro, weight, and samples averaging methods (Varapalli, 2022 ; Nkitgupta, 2021), Hamming loss (Ojeme and Mbogho, 2016). While precision, recall, and F1-score are some of the most common metrics in machine learning models, Hamming loss measures the hamming distance between labels predicted and true and gives penalties for wrongly classified instances for individual labels (3.4. Metrics and scoring: quantifying the quality of predictions, no date, sec. 3.4.4.8. Hamming loss).

As for the synthetic dataset, in total 142 samples were generated. The classification results are presented in the table (Table 1) below.

|  |  |  |
| --- | --- | --- |
| Classifier | Initial Dataset | Partially synthetic Dataset (MLSMOTE technique) |
| Classifier Chains (CC) with Random Forest - model 1 | Average Precision: 0.7706 Average Recall: 0.4634 Average F1 Score: 0.5206 Average Hamming Loss: 0.1130 | Average Precision: 0.9094 Average Recall: 0.8010 Average F1 Score: 0.8446 Average Hamming Loss: 0.0578 |
| Binary Relevance k-Nearest Neighbors (BRkNN) - model 2 | Average Precision: 0.4175 Average Recall: 0.1422 Average F1 Score: 0.2015 Average Hamming Loss: 0.1713 | Average Precision: 0.8633 Average Recall: 0.4751 Average F1 Score: 0.5883 Average Hamming Loss: 0.1435 |
| Multi-label k-Nearest Neighbors Classifier (MLkM) - model 3 | Average Precision: 0.3781 Average Recall: 0.2471 Average F1 Score: 0.2854 Average Hamming Loss: 0.1796 | Average Precision: 0.7636  Average Recall: 0.6120  Average F1 Score: 0.6699  Average Hamming Loss: 0.1375 |

*Table 1. Results for k-fold (n=5) cross-validation for 3 benchmarking classifiers*

Results suggest that with the classifiers work better with data augmentation. The average precision of the first model by 13.88%, for the second - 44,58%, and for the third - 38.55%. The positive changes in recall and f1-score are also noticeable, while the hamming loss was reduced as it should be with a better prediction outcome. Therefore, the other manipulations will be tested for the partially synthetic dataset.

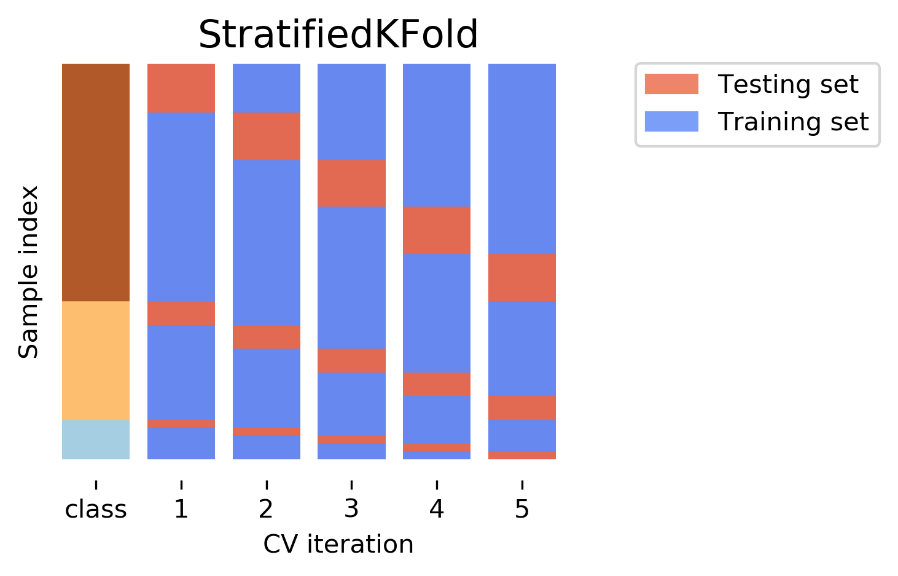
1. **Model development**

As previously mentioned, there is no established guideline for model selection for the multilabel classification problem. Therefore, all the candidate models previously tested for biomedical data were chosen along with others mentioned in the article. To evaluate the performance of the models, Hamming loss was selected as a primary metric along with others used to evaluate the semi-synthetic dataset. Following candidates for the classification algorithms were chosen: Classifier Chains (CC) with Random Forest (RF) and XGBoost base classifier, Binary Relevance and Random Forest base classifier, Random Forest (adapted for multi-label output), Multi-label k-NN, OneVsRest with XGBoost.

All implementations leveraged scikit-learn and XGBoost, Multi-label k-NN was implemented manually due to the problems with Python 3.11 version compatibility with the scikit-multilearn library. Since the grid search method did not support the hamming score evaluation automatically, the hamming loss scorer was implemented manually as well. The hamming loss was prioritized as it penalizes misclassifications per label, critical for clinical applications where overlooking any ADHD symptom can impact the ADHD-diagnosed type.

As for the other metrics used, the best-performing models were also evaluated with macro averages of precision, recall, and F1-score. Macro-averaging method is the most straightforward and easy-interpretable method, which does not assign more weights to any class (Karajgi, 2023).

Parameters in charge of control of a model’s architecture are called “hyperparameters”. A cross-validated grid search was applied for hyperparameter tuning to maximize model accuracy, this method is available through the GridSearchCV class of the scikit-learn module. The search utilized stratified k-fold cross-validation (k=5) to ensure that each target class had the same proportion as in the initial data. This technique addresses the issue of underrepresentation of the minority classes in folds to robust the generalization (Olamendy, 2024) (Figure 2.)



*Figure 2. Stratified K-fold cross validation, visualization of folds division.*

*Source:* [*https://www.kaggle.com/code/maheshkhanapure/cross-validation-and-types*](https://www.kaggle.com/code/maheshkhanapure/cross-validation-and-types)

1. **Feature importance analysis methodology**

To retrieve the importance of the features selected for each classifier, for all models except MLkNN, the built-in feature\_importance attribute (Gini importance for RF and gain-based for XGBoost) was used. Gini importance, or Mean Decrease Impurity, defined as the total decrease in node impurity weighted by the probability of reaching the node averaged for all trees (Lee, 2020), a popular choice for the ensemble learning techniques. Gain-based feature selection for XGBoost is a built-in algorithm, which measures the average gain of splits using the particular feature and indicates how much does it contribute to the performance (XGBooSt Best Feature Importance Score | XGBoosting, 20.04.2025). For multi-label models (Classifier Chains, Binary Relevance), importances were averaged across all label-specific classifiers, since they decompose the multi-label problem into multiple binary classifiers, but we need to identify globally useful features. For the MLkNN model, ANOVA F-values between each feature and the first label were computed, as k-NN lacks native importance metrics. This statistical method quantifies the relationships between features and labels.

The top five features were selected for the analysis of each model.

Results

The stratified 5-fold cross-validation on the MLSMOTE-augmented dataset revealed significant performance variations across models (Table 2).

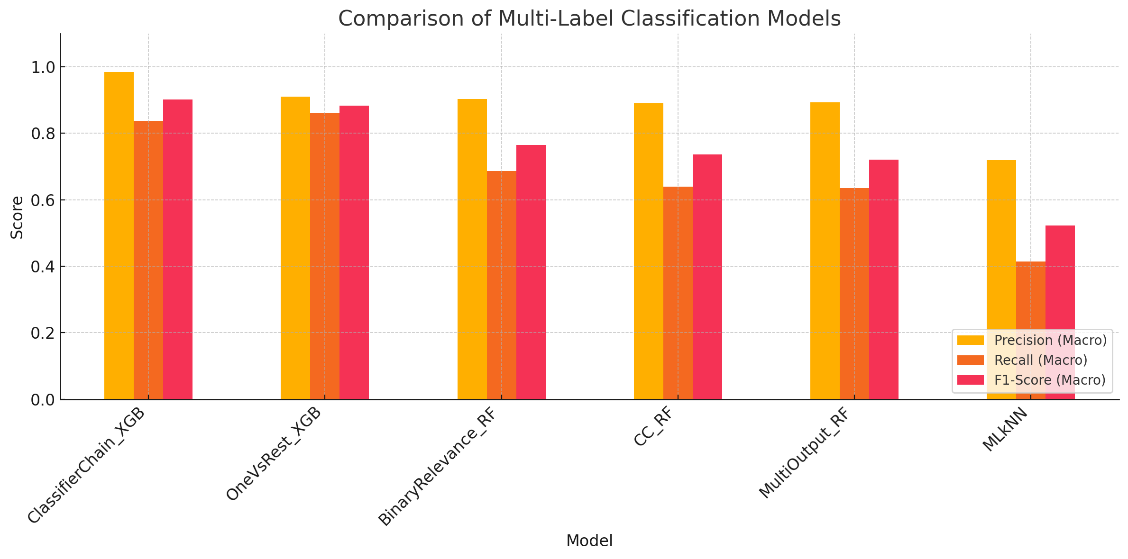
|  |  |  |
| --- | --- | --- |
| **Model** | **Best Hamming Loss** | **Key Parameters** |
| ClassifierChain\_XGB | 0.0741 | max\_depth=5, n\_estimators=100, lr=0.1 |
| OneVsRest\_XGB | 0.0898 | max\_depth=5, lr=0.1 |
| ClassifierChain\_RF | 0.1102 | max\_depth=10, n\_estimators=50 |
| BinaryRelevance\_RF | 0.1157 | max\_depth=10, n\_estimators=100 |
| MultiOutput\_RF | 0.1167 | max\_depth=10, n\_estimators=100 |
| MLkNN | 0.1861 | k=10 |

*Table 2. Hamming loss scores for the selected best-performing models for each type of classifiers.*

Based on the Hamming loss only the best model was a Classifier Chain with XGBoost classifier (Hamming loss 0.0741), demonstrating robust multi-label accuracy, while MLkNN performed weakest (Hamming loss=0.1861). The second-best approach was OneVsRest with XGBoost, which trailed closely (Hamming loss=0.0898).

It is noticeable, that Random-forest based models required deeper depth of the trees (max\_depth=10) thanthe XGBoost technique(max\_depth=5).

To adequately compare model performance, we calculated the macro averages of precision, recall, and F1-score metrics. Figure 3 below presents these metrics for each model.



*Figure 3. The scores comparison for models chosen based on the Hamming loss score.*

The classifier Chain with XGBoost maintained the highest macro precision (0.926), the highest macro recall (0.786), and f1-score (0.827), while the Classifier Chain with Random Forest and other tree-based models showed close macro precision scores. However, the recall and f1-score results were lower for other models, suggesting the selectiveness of the models with their prediction. This means that predictions are correct but the model does not predict every expected label, which results in a lower recall score (Karajgi, 2023).

The features with the top five values of importance were reported for each model and placed in Annex 1. Across all algorithms, no single feature consistently emerged as the dominant predictor, underscoring the variability in how different models prioritize symptom-related factors. Decision trees and ensemble methods frequently highlighted Question 11 or 12, while only two approaches—Classifier Chains with Random Forest and multiple-output k-NN—relied on composite scores (Part A and B). Notably, demographic features were absent from all top-five rankings, suggesting limited predictive utility in this context.

# Discussion & conclusions

In this work, it was demonstrated that machine learning algorithms can predict ADHD symptom profiles from self-reported survey data, with ensemble methods (XGBoost) leveraging classifier chains (CC) emerging as the most robust approach (Hamming loss: 0.074, F1-score: 0.827). The CC is a machine learning transformation technique, that constructs a Bayesian conditioned chain, where the first classifier is trained on the input space, and the next one is trained on both the input space and all previous classifiers of the chain (scikit-multilearn, 5.04.2025). In the discussion of Read et al. (2009), it was shown that the CC method is argued to overcome the disadvantages of Binary Relevance (BR) and achieve higher prediction accuracy by modeling label dependencies, which aligns with our results. Both CC implementations outperformed BR, suggesting meaningful symptom interdependencies in our data. This finding has clinical relevance, as the ASRS instrument is designed to classify ADHD subtypes, and the DSM-5 explicitly groups hyperactivity and impulsivity as a unified behavioral domain (American Psychiatric Association, 2013).

The second-best performance was observed with the One-vs-Rest (OvR) method using the same XGBoost classifier (Hamming loss: 0.0898, F1-score: 0.801). OvR is a heuristic strategy that decomposes the multi-label classification task into multiple independent binary classification problems, each focusing on a single label. While this method does not explicitly model label correlations, it nonetheless performed well, likely due to the strong generalization ability of the base learner, XGBoost.

XGBoost (Extreme Gradient Boosting) is a powerful and scalable implementation of gradient boosting algorithms. It builds additive models forward stage-wise, optimizing a differentiable loss function and incorporating both first- and second-order gradients for improved convergence. XGBoost is known for its ability to handle high-dimensional data, manage missing values internally, and provide regularization to reduce overfitting. These strengths make it especially suitable for tabular, structured data like survey responses. In this study, XGBoost likely compensated for the lack of explicit label interdependency modelling in OvR by effectively learning complex feature-label relationships, which supports its competitive performance despite the method’s simplifications.

Another finding is that the Random Forests model outperformed MLkNN. This may be due to MLkNN’s sensitivity to non-linear dependencies, also the choice of k and distance metric may have influenced the model outcome.

The divergence in top-ranked features across models (Annex 1) implies that feature importance is highly algorithm-dependent, complicating the identification of universal markers for ADHD aross the young group of people. The recurrent prominence of Questions 11/12 in tree-based models suggests these items may hold clinical relevance, whereas the exclusion of demographics aligns with prior findings on their weak discriminative power. While none of the accounted demographic variables were prioritized - consistent with the source study's findings showing weak correlation between background features and ADHD presence - other uncollected variables might be more relevant for preliminary ADHD diagnosis. For example, a study using deep neural networks identified parent criminal conviction, ADHD family history, educational failures, and speech/learning disabilities among the five most significant predictive features (Garcia-Argibay *et al.*, 2022b).

Several limitations of this study warrant discussion. First, the proportion of participants classified as ADHD-positive (17%) substantially exceeds the established adult ADHD prevalence of 3.1% (Ayano et al., 2023). This discrepancy likely reflects both the self-reported nature of our data and potential sampling bias, as individuals with suspected symptoms may have been more likely to participate. To mitigate overinterpretation, the groups were labelled as "ADHD-like" rather than clinical cases, emphasizing that the ASRS serves as a screening tool requiring diagnostic confirmation. Future studies should incorporate clinician-administered assessments to validate symptom classifications.

Second, while the models identified meaningful symptom patterns, the dataset lacked potentially informative covariates such as parental age, sibling ADHD status, or neurodivergent traits measurable by instruments like RAADS-R. Including these variables could improve subtype differentiation, particularly for subclinical presentations that often evade diagnosis.

Furthermore, the approach to label dominant symptoms—though necessary to address the low prevalence of ADHD subtypes in the dataset—deviated from DSM-5 guidelines, potentially introducing bias and reducing the dataset's representativeness. Specifically, the labelling protocol could have overemphasized certain symptoms that occur more frequently in this data sample but may not reflect the clinical population. While this symptom labelling approach achieved practical benefits for model training, future work should prioritize DSM-5-aligned classification schemes.

Methodologically, although I evaluated multiple algorithms, the lack of established guidelines for multi-label ADHD classification metrics complicates direct comparisons across studies. The use of MLSMOTE oversampling (Fernández et al., 2023) effectively balanced the labels but may have introduced synthetic patterns that require validation in naturalistic samples. The superior performance of XGBoost-based classifier chains indicates that ensemble methods better capture ADHD's symptom interdependencies, yet this requires replication for the fully real-world data.

While this study advances computational ADHD screening, its findings underscore the interplay between algorithmic choice, data completeness, and clinical validity. Future efforts must bridge these gaps to translate machine learning insights into robust diagnostic tools.

# Work evaluation

The first objective of this project—to develop a model for the prediction of ADHD symptom subtypes—was successfully achieved. However, the second objective could not be completed due to the unexpected complexity of the task. Specifically, this model required a more advanced approach to dataset preparation, including detailed symptom labelling and an additional step involving synthetic data generation and model evaluation.

An additional research question emerged during the course of the project: whether demographic data influences the model's predictions. This question was addressed as part of the work.

The time allocation for the project was somewhat adjusted due to the workload from other coursework, particularly the literature review. Nevertheless, the presentation was prepared and delivered on time.

One of the deliverables—the GitHub repository containing the project code—is still in progress and will be uploaded shortly.

# REFERENCES

Journal Articles:

Diakou, I., et al. (2024). Multi-label classification of biomedical data. Medicine International, 4(6). https://doi.org/10.3892/mi.2024.192

Folorunso, S. O., et al. (2020). A multi-label learning model for psychotic diseases in Nigeria. Informatics in Medicine Unlocked, 19, 100326. https://doi.org/10.1016/j.imu.2020.100326

Garcia-Argibay, M., et al. (2022b). Predicting childhood and adolescent attention-deficit/hyperactivity disorder onset: A nationwide deep learning approach. Molecular Psychiatry, 28(3), 1232-1239. https://doi.org/10.1038/s41380-022-01918-8

Gnanavel, S., et al. (2019). Attention deficit hyperactivity disorder and comorbidity: A review of literature. World Journal of Clinical Cases, 7(17), 2420-2426. https://doi.org/10.12998/wjcc.v7.i17.2420

Hines, J. L., King, T. S., & Curry, W. J. (2012). The Adult ADHD Self-Report Scale for screening for Adult Attention Deficit-Hyperactivity Disorder (ADHD). The Journal of the American Board of Family Medicine, 25(6), 847-853. https://doi.org/10.3122/jabfm.2012.06.120065

Karajgi, A. (2023, January 18). Evaluating multi-label classifiers - TDS Archive - Medium. Medium. https://medium.com/data-science/evaluating-multi-label-classifiers-a31be83da6ea

Magnus, W., Anilkumar, A. C., & Shaban, K. (2023). Attention deficit hyperactivity disorder. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK441838/

McCann, B. S., & Roy-Byrne, P. (2004). Screening and diagnostic utility of self-report attention deficit hyperactivity disorder scales in adults. Comprehensive Psychiatry, 45(3), 175-183. https://doi.org/10.1016/j.comppsych.2004.02.006

Morales, F. (2022, March 30). Multi-Class ML Model Evaluation - Analytics Vidhya - Medium. Medium. https://medium.com/analytics-vidhya/multi-class-ml-models-evaluation-103c9fdadb41

Ojeme, B., & Mbogho, A. (2016). Selecting learning algorithms for simultaneous identification of depression and comorbid disorders. Procedia Computer Science, 96, 1294-1303. https://doi.org/10.1016/j.procs.2016.08.174

Read, J., et al. (2009). Classifier chains for multi-label classification. In Lecture Notes in Computer Science (pp. 254-269). https://doi.org/10.1007/978-3-642-04174-7\_17

Rucklidge, J. J. (2010). Gender Differences in Attention-Deficit/Hyperactivity Disorder. Psychiatric Clinics of North America, 33(2), 357-373. https://doi.org/10.1016/j.psc.2010.01.006

Sadek, J. (2023). Attention Deficit Hyperactivity Disorder Misdiagnosis: Why medical evaluation should be a part of ADHD assessment. Brain Sciences, 13(11), 1522. https://doi.org/10.3390/brainsci13111522

Tarekegn, A. N., Giacobini, M., & Michalak, K. (2021). A review of methods for imbalanced multi-label classification. Pattern Recognition, 118, 107965. https://doi.org/10.1016/j.patcog.2021.107965

Van De Glind, G., et al. (2013). Validity of the Adult ADHD Self-Report Scale (ASRS) as a screener for adult ADHD in treatment seeking substance use disorder patients. Drug and Alcohol Dependence, 132(3), 587-596. https://doi.org/10.1016/j.drugalcdep.2013.04.010

Online Resources:

3.4. Metrics and scoring: quantifying the quality of predictions. (n.d.). scikit-learn. <https://scikit-learn.org/stable/modules/model_evaluation.html#hamming-loss>

ADHD in Adults: 4 Things to know. (2025, April 1). National Institute of Mental Health. <https://www.nimh.nih.gov/health/publications/adhd-what-you-need-to-know#part_6212>

Buchanan, B. (2025). Adult ADHD Self-Report Scale (ASRS). NovoPsych. <https://novopsych.com/assessments/diagnosis/adult-adhd-self-report-scale-asrs/>

Lee, C. (2020, September 8). Feature Importance Measures for Tree Models — Part I. \*Medium\*. <https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3>

Nkitgupta. (2021). Evaluation metrics for multi-class classification. Kaggle. <https://www.kaggle.com/code/nkitgupta/evaluation-metrics-for-multi-class-classification>

Olamendy, J. C. (2024, November 29). A comprehensive guide to Stratified K-Fold Cross-validation for unbalanced data. Medium. <https://medium.com/@juanc.olamendy/a-comprehensive-guide-to-stratified-k-fold-cross-validation-for-unbalanced-data-014691060f17>

scikit-multilearn. (n.d.). <http://scikit.ml/api/skmultilearn.problem_transform.cc.html>

Substance Abuse and Mental Health Services Administration. (2016). DSM-5 Changes: Implications for Child Serious Emotional Disturbance. NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK519712/table/ch3.t3/>

Varapalli, V. V. (2022, May 22). Evaluating metrics for Multi-label Classification and Implementations. Medium. <https://iamvishnu-varapally.medium.com/evaluating-metrics-for-multi-label-classification-and-implementations-ebabead60afd>

XGBooSt Best Feature Importance Score | XGBoosting. (2025, April 20). <https://xgboosting.com/xgboost-best-feature-importance-score/>

Annex

Annex 1. Table of top 5 feature importance for each of the chosen models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Feature Index** | **Feature Name** | **Importance Score** | **Question content:** |
| **CC\_RF** | **17** | Q11 | 0.100 | "Often leaves seat when remaining seated is expected" (Hyperactivity) |
|  | **7** | Q1 | 0.078 | "Often fails to give close attention to details" (Inattention) |
|  | 8 | Q2 | 0.070 | "Often has difficulty sustaining attention" (Inattention) |
|  | 15 | Q9 | 0.052 | "Often forgetful in daily activities" (Inattention) |
|  | 26 | 7-18 sum | 0.050 | Hyperactivity/Impulsivity composite score |
| **ClassifierChain\_XGB** | 7 | Q1 | 0.102 | "Often fails to give close attention" (Inattention) |
|  | 8 | Q2 | 0.090 | "Difficulty sustaining attention" (Inattention) |
|  | 17 | Q11 | 0.072 | "Leaves seat inappropriately" (Hyperactivity) |
|  | 18 | Q12 | 0.049 | "Often runs about/climbs excessively" (Hyperactivity) |
|  | 15 | Q9 | 0.042 | "Forgetful in daily activities" (Inattention) |
| **BinaryRelevance\_RF** | 11 | Q5 | 0.079 | "Difficulty organizing tasks" (Inattention) |
|  | 22 | Q16 | 0.078 | "Often blurts out answers" (Impulsivity) |
|  | 18 | Q12 | 0.071 | "Runs about/climbs excessively" (Hyperactivity) |
|  | 20 | Q14 | 0.053 | "Often 'on the go'" (Hyperactivity) |
|  | 17 | Q11 | 0.044 | "Leaves seat inappropriately" (Hyperactivity) |
| **MLkNN** (F-values) | 8 | Q2 | 41.34 | "Difficulty sustaining attention" (Inattention) |
|  | 17 | Q11 | 34.91 | "Leaves seat inappropriately" (Hyperactivity) |
|  | 7 | Q1 | 34.59 | "Fails to give close attention" (Inattention) |
|  | 25 | 1-6 sum | 28.74 | Inattention composite score |
|  | 15 | Q9 | 24.10 | "Forgetful in daily activities" (Inattention) |
| **MultiOutput\_RF** | 18 | Q12 | 0.077 | "Runs about/climbs excessively" (Hyperactivity) |
|  | 22 | Q16 | 0.075 | "Blurts out answers" (Impulsivity) |
|  | 11 | Q5 | 0.071 | "Difficulty organizing tasks" (Inattention) |
|  | 20 | Q14 | 0.059 | "Often 'on the go'" (Hyperactivity) |
|  | 7 | Q1 | 0.043 | "Fails to give close attention" (Inattention) |
| **OneVsRest\_XGB** | 11 | Q5 | 0.080 | "Difficulty organizing tasks" (Inattention) |
|  | 18 | Q12 | 0.076 | "Runs about/climbs excessively" (Hyperactivity) |
|  | 12 | Q6 | 0.062 | "Often avoids tasks requiring sustained mental effort” (Hyperactivity (Verbal) **)** |
|  | 22 | Q16 | 0.059 | "Often blurts out answers before questions completed" (Impulsivity) |
|  | 21 | Q15 | 0.058 | "Often talks excessively”m |