國立臺灣大學電機資訊學院生醫電子與資訊學研究所

博士論文

Graduate Institute of Biomedical Electronics and Bioinformatics
College of Electrical Engineering and Computer Science
National Taiwan University

Doctoral Dissertation

運用機器學習方法檢測小兒發展遲緩病症

Detection of Pediatric Developmental Delay (DD) with Machine Learning Technologies

陳 新 博

Shin-Bo Chen

指導教授：歐陽彥正 博士

Advisor: Yen-Jen Oyang, Ph.D.

中華民國 114 年 9 月

Sep. 2025

doi:10.6342/NTU202504505

國立臺灣大學博士學位論文
口試委員會審定書
PhD DISSERTATION ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

運用機器學習方法檢測小兒發展遲緩病症

Detection of Pediatric Developmental Delay (DD) with Machine Learning Technologies

本論文係 **陳新博** D00945012 在國立臺灣大學電機資訊學院生醫電子與資訊學研究所完成之博士學位論文，於民國 114 年 09 月 22 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Biomedical Electronics and Bioinformatics on 22/09/2025 have examined a PhD dissertation entitled above presented by SHIN-BO CHEN D00945012 candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

歐陽彥正	傅楸善	張瑞峰
（指導教授 Advisor）		
李子相	楊孟翰	

系主任/所長 Director: 林風

doi:10.6342/NTU202504505

誌謝

回首這段漫長而深邃的求學旅程，心中湧起無限感懷。自入學以來，歷經多次休學與再啟程的歲月，在工作責任與家庭照護之間奔波往返，常於深夜與黎明交會之時，獨自思索研究的方向與意義。雖然前行的步伐緩慢，卻從未停歇。這些年寒窗苦讀的歷程，讓我深切體悟「堅持」並非僅是一種行動，而是一種信念、一份靜默而持久的力量。如今能夠完成博士論文，並將研究成果發表於國際期刊，所有的辛勞都化為無以言喻的感動與平靜的喜悅。

衷心感謝指導教授 歐陽彥正 教授 在整個研究歷程中的悉心指導與啟迪。老師以嚴謹的學術態度與深厚的洞見，引領我從研究設計、資料分析到論文撰寫的每一環節；更以開放與信任的胸懷，讓我能從臨床實務出發，探索醫療現場真實的問題，並結合博士班所學之機器學習專業，發展出兼具理論深度與實務價值的研究成果。老師的睿智、包容與信任，是我能走到今日最堅實的依靠與力量。

誠摯感謝一路陪伴與支持我的家人與同事。家人的理解、體諒與溫柔包容，讓我能在繁忙與疲憊之間仍保持專注；同事與朋友們於臨床資料協助、技術支援與精神鼓勵上的相伴，使我在無數難關中依然能心懷熱忱。你們的存在，是我得以堅持的理由與溫度。

更要感謝那個不曾放棄的自己—在無數個深夜裡，仍堅持完成每一段程式、修訂每一頁文字的自己。這份學位，不僅是學術旅程的終點，更是人生修行的一部分；它見證了信念、耐心與成長，也成就了今日的我。

謹以此文，獻給所有在我生命中給予理解、支持與啟發的人。

doi:10.6342/NTU202504505

Acknowledgements

This dissertation concludes a long and challenging journey. Balancing study, professional duties, and family responsibilities through years of interruption and return has taught me the true meaning of perseverance. Completing this work and seeing it published in an international journal fills me with deep gratitude and quiet joy.

I am sincerely grateful to Professor Yen-Jen Oyang for his insightful guidance, academic rigor, and generous trust. His mentorship encouraged me to draw from real clinical experiences and apply machine learning methods acquired in the Ph.D. program to develop practical solutions bridging theory and practice.

I also wish to thank my family and colleagues for their patience, understanding, and constant support. Their faith and encouragement sustained me through every challenge.

This accomplishment stands as both an academic milestone and a personal journey of endurance, reflection, and growth.

doi:10.6342/NTU202504505

中文摘要

運用機器學習方法檢測小兒發展遲緩病症

研究目的

對臨床治療師而言，準確鑑別可能出現發展遲緩（DD）的兒童始終是一項挑戰。近年研究指出，若兒童能及早接受介入治療，其臨床預後顯著優於未接受介入者。本研究旨在探討兒童接受三類治療（物理治療、職能治療與語言治療）的頻率，是否可作為檢測其是否罹患發展遲緩的依據。此方法的核心價值在於，相關特徵取得成本極低，若能建立有效的預測模型，將可用於初步篩檢，在進行昂貴且複雜的診斷程序之前，先行辨識可能有 DD 風險的兒童。

研究方法

本研究使用台灣某醫院於 2012 至 2016 年間蒐集之臨床資料，共涵蓋 2,552 位門診個案（共 34,862 筆就診紀錄，平均年齡 72.34 月）。基於該資料集，本研究分別建立三種機器學習預測模型：深度神經網路（Deep Neural Network, DNN）、支援向量機（Support Vector Machine, SVM）以及決策樹（Decision Tree, DT），以評估所提出方法的效能。

研究結果

實驗結果顯示，就 F1 分數（靈敏度與陽性預測值的調和平均數）而言，當需要維持高靈敏度時，DT 模型的表現優於 DNN 與 SVM 模型。具體而言，本研究所建立的 DT 模型達成靈敏度 0.902 與陽性預測值 0.723 的表現。

研究結論

本研究結果顯示，兒童接受治療的頻率蘊含重要訊息，能有效應用於發展遲緩的預測。由於

doi:10.6342/NTU202504505

此類特徵可在不增加額外成本的情況下獲取，且實驗結果展現良好效能，因此可合理推測，依此建立之預測模型具備高度臨床應用潛力，並有望顯著改善發展遲緩兒童的治療成效。

關鍵字：發展遲緩、職能治療服務、治療頻率、機器學習、決策樹、支援向量機、深度神經網路

doi:10.6342/NTU202504505

Abstract

Detection of Pediatric Developmental Delay (DD) with Machine Learning Technologies

Objective

Accurate identification of children who will develop delay (DD) is challenging for therapists because recent studies have reported that children who underwent early intervention achieved more favorable outcomes than those who did not. In this study, we have investigated how the frequencies of three types of therapy, namely the physical therapy, the occupational therapy, and the speech therapy, received by a child can be exploited to predict whether the child suffers from DD or not. The effectiveness of the proposed approach is of high interest as these features can be obtained with essentially no cost and therefore a prediction model built accordingly can be employed to screen the subjects who may develop DD before advanced and costly diagnoses are carried out.

Methods

This study has been conducted based on a data set comprising the records of 2,552 outpatients (N = 34,862 visits, mean age = 72.34 months) collected at a hospital in Taiwan from 2012 to 2016. We then built 3 types of machine learning based prediction models, namely the deep neural network

doi:10.6342/NTU202504505

models (DNN), the support vector machine (SVM) models, and the decision tree (DT) models, to evaluate the effectiveness of the proposed approach.

Results

Experimental results reveal that in terms of the F1 score, which is the harmonic mean of the sensitivity and the positive predictive value, the DT models outperformed the DNN models and the SVM models, if a high level of sensitivity is desired. In particular, the DT model developed in this study delivered the sensitivity at 0.902 and the positive predictive value at 0.723.

Conclusions

What has been learned from this study is that the frequencies of the therapies that a child has received provide valuable information for predicting whether the child suffers from DD. Due to the performance observed in the experiments and the fact that these features can be obtained essentially without any cost, it is conceivable that the prediction models built accordingly can be wide exploited in clinical practices and significantly improve the treatment outcomes of the children who develop DD.

Keywords: Developmental Delay, Occupational Therapy service, Frequency of therapy, Machine Learning, Decision Tree, Support Vector Machine, Deep Neural Networks

doi:10.6342/NTU202504505

目次

doi:10.6342/NTU202504505

doi:10.6342/NTU202504505

List of Figures

doi:10.6342/NTU202504505

List of Tables

doi:10.6342/NTU202504505

Chapter I Introduction

1-1 Background

Developmental delay (DD) refers to a distinct set of early childhood developmental disabilities, and it is primarily diagnosed by assessing a child's behavioral and mental capacities [1]. Rehabilitation physicians and pediatric specialists employ a wide range of diagnostic strategies, clinical tools, and classification frameworks to assess and manage developmental delay (DD). These approaches encompass structured physical and neurological examinations, standardized motor and cognitive assessments, sensory evaluations (such as hearing and vision screening), genetic and metabolic testing, and neuroimaging modalities, all of which are essential for identifying underlying etiologies and guiding therapeutic planning [2-5]. Furthermore, systematic classification systems and rehabilitation taxonomies are increasingly integrated into clinical decision-making to stratify patients, personalize intervention strategies, and predict therapeutic outcomes, thereby enhancing the precision and effectiveness of early intervention programs [6]. However, these classification methods are often subjective, time-consuming, and prone to inconclusive results. Moreover, they fail to clarify the underlying causes and are ineffective for early detection [7]. Early intervention significantly improves a child's likelihood of reaching their full potential, with studies reporting that children who

doi:10.6342/NTU202504505

receive early intervention achieve more favorable outcomes than those who do not. [8,9-13] Therefore, accurate classification of DD is crucial for providing effective early intervention services that ensure positive outcomes for children with DD.

Machine learning has been employed to develop novel computational methods that incorporate mathematical learning, statistical estimation, and information theories [14]. These methods automatically identify meaningful patterns within large datasets. A key advantage of machine learning is its ability to generate highly accurate and reliable predictions based on data comprising multiple variables. Additionally, machine learning enables causal inference from non-experimental datasets [15].

In recent years, machine learning (ML) has emerged as a powerful tool in psychiatric and neurodevelopmental research, demonstrating its potential to detect and classify complex conditions such as autism spectrum disorder (ASD) [16], attention deficit hyperactivity disorder (ADHD) [17], and schizophrenia [18]. By extracting subtle patterns from high-dimensional data, ML has been shown to surpass conventional statistical methods in diagnostic accuracy and predictive performance. For example, Bishop et al. successfully applied ML to predict the lifetime health trajectories of adults with ASD, accurately forecasting the onset of cardiovascular, urinary, and respiratory

doi:10.6342/NTU202504505

conditions [19]. These advances have positioned ML as a promising technology for early detection and personalized healthcare.

Despite these achievements, existing ML approaches suffer from critical limitations that restrict their clinical applicability—particularly in the context of early intervention for developmental delay (DD). Most prior studies rely on costly and time-consuming data sources, such as neuroimaging, electrophysiological signals, or extensive behavioral assessments. These methods require specialized equipment, expert personnel, and complex analytical workflows, making them impractical for large-scale implementation in community or primary care settings. Moreover, the heavy dependence on retrospective datasets and resource-intensive pipelines often results in delayed outputs, limiting their utility for pre-symptomatic screening and proactive intervention. Consequently, although traditional ML techniques have demonstrated strong predictive capabilities, their lack of scalability, high operational costs, and lengthy processing times hinder their adoption in real-world clinical workflows, particularly in scenarios that demand timely decision-making and broad population coverage.

1-2 Motivation

doi:10.6342/NTU202504505

This gap is particularly concerning given the well-established importance of early intervention in developmental disorders. Studies have reported that most cases of DD gradually resolve over time [20], underscoring the value of identifying at-risk children as early as possible and providing targeted support before long-term functional deficits emerge. However, few studies to date have explored how ML can be harnessed not only to classify DD but also to identify key predictive factors that inform the timing, intensity, and optimization of intervention strategies. There remains a significant unmet need for predictive models that are cost-effective, scalable, and capable of integrating seamlessly into routine healthcare delivery.

The novelty of our approach lies in addressing this gap by proposing the use of therapy utilization—specifically, the frequency of occupational, physical, and speech therapy services—as a predictive factor for DD classification and outcome forecasting [21-29]. Therapy frequency is a clinically meaningful and readily available variable that reflects both the severity of developmental challenges and the child's responsiveness to intervention. Unlike imaging or specialized assessments, therapy utilization data are routinely collected in clinical practice, require no additional cost or infrastructure, and can be easily incorporated into electronic health records. By integrating such features

doi:10.6342/NTU202504505

into ML models, we aim to develop predictive systems that are not only accurate but also practical and scalable.

This approach offers several important advantages. First, it leverages existing clinical data to enable large-scale population screening without the financial and logistical barriers associated with traditional ML models. Second, it supports real-time decision-making by allowing clinicians to monitor therapy responsiveness and adjust intervention strategies dynamically. Third, it enhances personalization in clinical care by identifying individual-level predictors of developmental outcomes, thereby facilitating the design of tailored therapeutic plans. Ultimately, by bridging the gap between algorithmic sophistication and clinical feasibility, our study seeks to transform ML from a research tool into a practical, actionable framework for early intervention in developmental delay. This paradigm shift holds the potential to improve developmental trajectories, reduce healthcare burdens, and ensure that at-risk children receive the right support at the right time.

1-3 Organization of this thesis

This dissertation is organized into seven chapters, each designed to build a logical and coherent narrative from foundational motivation to methodological development,

doi:10.6342/NTU202504505

empirical validation, and future research directions. The structure reflects the interdisciplinary nature of this work, which integrates clinical knowledge, machine learning techniques, and public health considerations to advance early detection strategies for pediatric developmental delay (DD).

Chapter I – Introduction establishes the conceptual foundation of the study. It presents the clinical and societal significance of early detection and intervention in DD, identifies critical limitations of existing diagnostic approaches, and articulates the central research questions and objectives. The chapter also outlines the study's hypotheses and theoretical framework, situating the research within the broader context of computational medicine and predictive healthcare.

Chapter II – Literature Review surveys the state of the art in machine learning applications for neurodevelopmental disorder classification, including autism spectrum disorder (ASD) and other developmental conditions. It critically analyzes a range of existing approaches—from classical classifiers such as linear discriminant analysis (LDA) and K-nearest neighbors (K-NN) to ensemble learning, deep learning, and multimodal fusion frameworks—and compares their performance, methodological assumptions, and clinical applicability. Special attention is given to the persistent

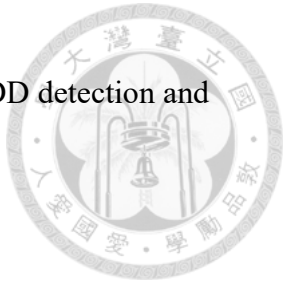
doi:10.6342/NTU202504505

challenges of high cost, lengthy data acquisition, and limited scalability, which collectively underscore the necessity for innovative, cost-efficient approaches such as the one proposed in this dissertation.

Chapter III – Methods describes the research design, dataset characteristics, and feature engineering strategies employed in this study. It explains the selection of key variables, including therapy utilization frequency (occupational, physical, and speech therapy sessions), and details the preprocessing steps, feature selection techniques, and modeling pipeline. The chapter also elaborates on the implementation of three predictive models—Decision Tree (DT), Support Vector Machine (SVM), and Deep Neural Network (DNN)—and outlines the evaluation metrics and validation procedures, including cross-validation, receiver operating characteristic (ROC) analysis, and precision-recall assessments.

Chapter IV – Results presents the experimental findings and performance outcomes of the developed models. It provides a comprehensive analysis of predictive accuracy, sensitivity, specificity, precision, and F1-scores, as well as graphical representations such as ROC and precision-recall curves. The results are interpreted in relation to model

doi:10.6342/NTU202504505



robustness, generalizability, and their clinical implications for early DD detection and patient stratification.

Chapter V – Discussion offers an in-depth interpretation of the results and situates them within the context of existing literature and clinical practice. It discusses the advantages of therapy-based predictive modeling relative to traditional approaches, the implications for early intervention strategies, and the potential for integration into healthcare workflows. The chapter also addresses methodological considerations, practical deployment challenges, and the broader impact of predictive analytics on personalized rehabilitation planning.

Chapter VI – Conclusion synthesizes the key contributions and findings of the dissertation. It emphasizes the novelty of leveraging therapy utilization as a predictive feature, the methodological advancements introduced, and the clinical value of developing a low-cost, scalable, and interpretable ML framework for early DD detection. The chapter further reflects on the study's potential to influence future screening protocols and public health interventions.

Chapter VII – Future Works outlines potential avenues for extending this research. These include exploring ensemble and hybrid modeling strategies, incorporating

doi:10.6342/NTU202504505

additional clinical and sociodemographic features, validating models across larger and more diverse populations, and developing longitudinal predictive systems. The chapter also addresses ethical, practical, and translational considerations that will be essential for advancing predictive models toward clinical deployment.

doi:10.6342/NTU202504505

Chapter II Literature Reviews

Machine learning (ML) has become an increasingly powerful tool in the field of neurodevelopmental and psychiatric research, offering new ways to analyze complex datasets and identify subtle patterns that may underlie developmental disorders such as autism spectrum disorder (ASD) and developmental delay (DD). Researchers have explored a wide range of classification algorithms to diagnose and predict these conditions, each contributing to the evolving landscape of computational neurodevelopmental science. Among the earliest methods employed were linear discriminant analysis (LDA) and K-nearest neighbors (K-NN), both of which were used to classify ASD based on behavioral and questionnaire data [21]. Osman Altay and colleagues reported that LDA outperformed K-NN in terms of precision, highlighting the importance of algorithm selection in improving classification performance. Similarly, Fatiha Nur and Ali Öztürk compared several classifiers—including random forest (RF), naïve Bayes (NB), K-NN, and radial basis function networks (RBFN)—and concluded that RF achieved superior predictive outcomes [22]. These studies illustrate the foundational role of traditional classification algorithms in early ASD research and demonstrate how algorithmic choices can significantly affect diagnostic accuracy.

doi:10.6342/NTU202504505

As ML techniques evolved, the integration of neuroimaging data into classification models marked a significant methodological advancement. Imaging modalities such as structural magnetic resonance imaging (sMRI) and resting-state functional MRI (rs-fMRI) offer direct insights into neural connectivity, cortical structure, and brain organization, thereby enabling more biologically grounded models of DD classification. Dvornek et al. were among the first to combine phenotypic data with rs-fMRI using deep learning, achieving improved classification accuracy for ASD compared with traditional machine learning techniques [23]. Liao et al. proposed an innovative approach that incorporated community structure analysis with deep learning models, further enhancing predictive performance [24]. Dekhil et al. advanced this work by integrating anatomical and functional information from sMRI and fMRI, successfully distinguishing between autism and typical development [25]. These studies collectively demonstrate the substantial benefits of combining brain imaging data with ML techniques, suggesting that neurobiological features provide critical discriminative power for developmental disorder classification.

Beyond anatomical data, researchers have also focused on cortical measures and functional connectivity patterns as important predictors of DD. Surface-based morphometry (SBM) approaches, such as those used by Yun Jiao and colleagues,

doi:10.6342/NTU202504505

revealed cortical thickness to be a key predictive feature for ASD classification [26]. This finding underscores the potential of structural brain features as diagnostic biomarkers. In parallel, Heinsfeld et al. examined functional communication patterns derived from brain imaging and identified neural structures that differed significantly between ASD and typically developing individuals [27]. These contributions highlight how combining structural and functional information deepens our understanding of the neurobiological basis of DD, offering a more comprehensive view of how brain organization correlates with developmental outcomes. Such insights support the development of diagnostic models that are not only statistically robust but also biologically meaningful.

More recent advances have focused on the refinement of classification algorithms to improve both sensitivity and specificity, particularly in clinical settings where diagnostic precision is paramount. Bone et al. developed highly adaptable algorithms that outperformed existing methods by enabling separate optimization of sensitivity and specificity [28]. Their models, which analyzed data from standardized diagnostic instruments such as the Autism Diagnostic Interview-Revised (ADI-R) and the Social Responsiveness Scale (SRS), demonstrated strong potential for clinical application. Similarly, Jin et al. applied multi-kernel support vector machine (SVM) methods to

doi:10.6342/NTU202504505

classify infants at high risk for ASD as early as six months of age, integrating features related to white matter tracts and whole-brain connectivity [29]. Their approach outperformed single-scale network models, demonstrating the feasibility of early risk prediction using ML. Expanding beyond ASD, Kim et al. reported that SVM could outperform conventional statistical models in predicting the prognosis of Class III malocclusion, demonstrating the broader applicability of ML techniques beyond neurodevelopmental conditions [30]. Collectively, these advancements reflect significant methodological progress, showcasing how improvements in model architecture and feature engineering can translate into clinically relevant diagnostic tools.

However, despite these promising developments, the literature reveals several critical limitations that hinder the widespread clinical adoption of existing ML-based DD classification approaches. A recurring challenge is the heavy reliance on diagnostic symptom data and specialized biomarkers. Many models are built on features derived from neuroimaging, standardized behavioral scales, or electrophysiological signals—data sources that are often expensive, time-consuming, and difficult to obtain in routine clinical practice. For instance, MRI and fMRI acquisitions require costly infrastructure, trained personnel, and complex preprocessing pipelines, all of which contribute to

doi:10.6342/NTU202504505

substantial financial and logistical barriers. Moreover, such studies typically involve small, highly controlled samples, raising concerns about their generalizability to broader populations. This dependency on specialized diagnostic data can significantly limit the scalability and practicality of ML models, particularly in community-based healthcare settings where resources are constrained and large-scale screening is necessary.

Another major limitation concerns the temporal dimension of data acquisition. Most existing ML models rely on retrospective datasets, which means that predictions are often made after clinical symptoms have already emerged. As a result, these models are less effective in supporting proactive decision-making or early intervention strategies, which are crucial for improving long-term developmental outcomes. The time-intensive nature of data collection—especially in neuroimaging-based studies—also delays the deployment of ML tools in real-world settings. This is a significant drawback given that early intervention has consistently been shown to improve developmental trajectories and functional outcomes in children with DD. The gap between the timing of symptom onset and the availability of predictive information remains one of the most pressing challenges in the field.

doi:10.6342/NTU202504505

Data availability and quality also pose significant challenges. Because many ML models depend on high-quality imaging or diagnostic data, they may not perform well in real-world clinical environments where data are heterogeneous, incomplete, or noisy. Moreover, behavioral assessments often rely on subjective caregiver reports, which can introduce cultural and reporting biases. These limitations collectively reduce the reliability, reproducibility, and clinical utility of many existing ML-based approaches [31–34]. Even when high classification accuracy is achieved in controlled research settings, translating these models into clinical workflows remains difficult due to cost, complexity, and data availability constraints.

A deeper examination of previous research further reveals that methodological differences among machine learning approaches significantly influence their clinical applicability and predictive power. For instance, classical algorithms such as LDA and K-NN [21] laid the groundwork for early classification efforts by demonstrating the feasibility of computational diagnosis based on behavioral data. However, their relatively limited capacity to capture nonlinear patterns and complex feature interactions restricted their predictive accuracy. Ensemble methods such as random forests (RF) [22] improved upon these limitations by integrating multiple decision trees and leveraging feature importance measures, leading to more robust classification outcomes.

doi:10.6342/NTU202504505

Nonetheless, even these methods remained constrained by their dependence on well-curated datasets and their inability to dynamically adapt to heterogeneous, real-world clinical inputs.

The integration of neuroimaging into ML models [23–25] represented a major methodological leap, offering biologically grounded insights and improving classification precision. Dvornek et al. [23] demonstrated that the inclusion of rs-fMRI data with phenotypic variables significantly enhanced the discriminative performance of deep learning models, illustrating the power of multimodal feature fusion. Liao et al. [24] expanded on this approach by using community structure analysis to better capture the topological organization of neural networks, yielding superior accuracy compared to conventional classifiers. Dekhil et al. [25] further advanced this paradigm by combining sMRI and fMRI data, demonstrating that structural and functional features jointly provide a richer and more nuanced representation of developmental disorders. These contributions underscore the potential of imaging-based features to elucidate the neurobiological underpinnings of ASD and DD. However, they also illustrate a critical trade-off: as model complexity and predictive accuracy increase, so too do the cost, computational requirements, and barriers to clinical implementation.

doi:10.6342/NTU202504505

Further methodological innovations explored structural and functional metrics derived from cortical measures and brain communication patterns [26,27]. Surface-based morphometry studies, such as those conducted by Jiao et al. [26], identified cortical thickness as a key biomarker, while functional connectivity analyses by Heinsfeld et al. [27] provided insights into altered network dynamics in ASD populations. These studies not only advanced the field's understanding of neurodevelopmental pathology but also demonstrated that ML can effectively detect subtle, spatially distributed neural differences. Yet, the need for high-resolution imaging data and specialized analytical workflows remains a significant limitation, reducing the feasibility of deploying such models outside of well-equipped research institutions.

The development of more sophisticated algorithms aimed to overcome some of these limitations by increasing diagnostic sensitivity and specificity. Bone et al. [28] designed algorithms capable of weighting sensitivity and specificity independently, which is particularly valuable for clinical contexts where false negatives or false positives carry significant consequences. Jin et al. [29] extended these advances by demonstrating that multi-kernel SVM could classify high-risk infants as early as six months, thereby showing that ML has the potential to identify developmental vulnerabilities long before traditional diagnostic criteria can be applied. Kim et al. [30] further confirmed the

doi:10.6342/NTU202504505

clinical utility of SVM in non-neurodevelopmental contexts, such as orthodontic prognosis, reinforcing the adaptability of these methods to a range of predictive healthcare applications. Despite these advances, the cost and time required to collect and process imaging data, coupled with the limited availability of longitudinal datasets, continue to impede large-scale clinical adoption.

A comparative synthesis of the studies summarized in **Table 1** illustrates both the progress and the persistent challenges in ML-based DD classification. Classical classifiers (e.g., LDA, K-NN) demonstrated feasibility but were limited in handling complex data [21]. Ensemble methods (RF, GBM) improved predictive power but still required carefully engineered features [22]. Deep learning approaches significantly enhanced classification performance when combined with phenotypic and imaging data [23,24], while multimodal integration yielded the highest accuracies by leveraging complementary structural and functional features [25]. Surface-based morphometry and connectivity analyses enriched the interpretability of models and deepened the understanding of disease mechanisms [26,27]. More advanced classifiers, such as multi-kernel SVM and specialized ensemble algorithms, demonstrated potential for early detection and high-risk screening [28,29]. Yet, across all these methodologies, a common challenge persists: the reliance on expensive, specialized, and time-consuming

doi:10.6342/NTU202504505

data collection processes, as well as the limited scalability of models to real-world healthcare settings.

Table 1. A summary of the existing machine learning based predictors for identifying patients who may develop DD.

Classifier Used	Modality	Number of Subjects	Features	Performance	Contribution / Key Insight
RF, GBM [7]	MRI	876 (417 ASD, 459 TD)	White matter, gray matter, CSF, total intracranial volume	RF ACC = 60%	Pioneered the use of neuroanatomical MRI features to classify developmental delay, establishing early evidence for the potential of brain structural biomarkers despite limited accuracy.
LDA, K-NN [21]	Questionnaire	292 (141 ASD, 151 Non-ASD, 4–11 Yrs)	19 behavioral attributes/questions	K-NN **AUC** = 61%	Demonstrated the feasibility of low-cost, questionnaire-based screening, revealing early-stage predictive potential and emphasizing the need for richer feature sets.
NB, K-NN, RBFN, RF [22]	Questionnaire	244	21 behavioral attributes/questions	RF ACC = 96.4%	Highlighted the significant impact of classifier selection, showing that ensemble methods like RF enhance specificity even with non-clinical behavioral data.
NN [23]	rsfMRI	1100 (529 ASD, 571 controls)	Phenotypic features (age, sex, handedness, IQ, eye status)	DNN ACC = 70.1%	Combined resting-state fMRI and phenotypic data to improve classification accuracy, illustrating the added value of integrating neuroimaging with demographic variables.
Deep Learning [24]	rsfMRI	Group I: 38; Group II: 110; Group III: 35	NMI matrix, Pearson correlation matrix	NMI ACC = 59.09%	Advanced the application of deep learning to correlation-based connectivity features, highlighting both the promise and challenges of dataset variability.
MDN [25]	sMRI, fMRI	47 (22 ASD, 25 controls)	Cerebral cortex, white matter volumes	MDN ACC = 94.7%	Introduced a multimodal fusion approach combining structural and functional MRI, significantly enhancing performance and demonstrating complementary feature synergy.
SVM, FT, LMT [26]	MRI	38 (22 ASD, 16 controls)	Cortical thickness, curvature metrics, folding indices	FT/LMT ACC = 76%	Employed detailed cortical morphometry to improve model interpretability and classification accuracy, contributing to refined neuroimaging feature engineering.
DNN, SVM, RF [27]	rsfMRI, sMRI	1035 (505 ASD, 530 controls)	Phenotypic features (age, sex, handedness, IQ, eye status)	DNN ACC = 70%	Demonstrated the superiority of deep learning over classical models for complex, multi-modal feature sets, underscoring the value of data integration.
MLCV, SVM [28]	Questionnaire	1726 (1264 ASD, 462 non-ASD)	Correlation-based features	MLCV ACC = 89.2%	Achieved high predictive accuracy using advanced correlation-based feature engineering, proving that well-designed feature extraction can offset modality limitations.
SVM [29]	MRI	80 (40 high-risk infants, 40 low-risk)	Multiscale connectivity network	SVM ACC = 76%	Focused on early detection in high-risk infants, revealing the potential of functional connectivity markers for pre-symptomatic identification of developmental risk.

Abbreviations: linear discriminant analysis (LDA), mixture density network (MDN), naïve bayes (NB), k-nearest neighbor (KNN), radial basis function (RBF), gradient boosting model (GBM), normalized mutual information (NMI), resting-state functional magnetic resonance imaging (rsfMRI), structural MRI (sMRI), functional MRI (fMRI), deep neural network (DNN), random forest (RF), functional tree (FT), machine learning (ML), logistic model tree (LMT), accuracy (ACC) and area under curve (AUC).

The limitations of existing approaches have important implications for early intervention, which remains the most effective strategy for improving long-term outcomes in children with developmental delays. Early intervention can significantly

doi:10.6342/NTU202504505

enhance cognitive, social, and behavioral development, but its success depends on the timely identification of children at risk. Unfortunately, the reliance on post-symptomatic diagnostic data, costly imaging techniques, and complex modeling workflows delays the point of identification, reducing the window of opportunity for early support. Furthermore, the resource-intensive nature of these approaches restricts their deployment to specialized research or tertiary care centers, leaving many children—particularly those in underserved regions—without access to early diagnostic services. These systemic barriers highlight the urgent need for predictive models that are not only accurate but also cost-effective, scalable, and easily integrated into routine clinical workflows.

It is within this context that the present study proposes an innovative approach centered on therapy utilization—specifically, the frequency of physical, occupational, and speech therapy sessions—as a predictive feature for DD classification and outcome prediction. This paradigm shift addresses several of the key limitations identified in the literature. First, therapy frequency is a readily available and routinely collected clinical variable, requiring no specialized equipment, imaging resources, or additional financial investment. Second, because therapy utilization directly reflects both the severity of a child's developmental condition and their responsiveness to intervention, it serves as a

doi:10.6342/NTU202504505

meaningful proxy for underlying developmental trajectories. Third, leveraging such data allows predictive models to be applied across diverse healthcare settings, including primary care and community-based programs, thereby increasing their accessibility and public health impact.

The potential advantages of this approach extend beyond scalability and cost-effectiveness. By transforming therapy frequency into a predictive feature, ML models can not only identify at-risk children earlier but also guide personalized treatment planning. For example, patterns of therapy utilization may reveal which children are likely to respond to standard intervention protocols and which may require more intensive or specialized support. Additionally, because therapy frequency is a modifiable variable, predictive models based on this feature have the potential to inform real-time clinical decision-making—enabling practitioners to adjust therapy plans dynamically in response to predicted outcomes. This feedback loop between prediction and intervention represents a significant advancement over existing ML frameworks, which are primarily diagnostic rather than prescriptive in nature.

In summary, a comprehensive review of the literature reveals significant progress in the application of machine learning to the classification and prediction of developmental

doi:10.6342/NTU202504505

disorders. Studies employing a variety of algorithms—from classical classifiers to deep learning and multimodal approaches—have demonstrated the feasibility and potential of ML in enhancing diagnostic precision [21-30]. However, the widespread adoption of these methods remains constrained by high costs, time-consuming data acquisition processes, and limited scalability [31-34]. These challenges underscore the need for innovative solutions that bridge the gap between methodological sophistication and clinical feasibility. The present study responds to this need by proposing a novel, low-cost, and scalable predictive framework based on therapy utilization data. By exploiting information that is already collected in routine clinical practice, this approach offers a practical pathway toward early detection, individualized intervention planning, and improved developmental outcomes. Ultimately, this strategy holds the potential to transform ML-based DD classification from a research-oriented endeavor into a widely deployable clinical tool, enabling more timely and effective support for children at risk of developmental delay.

doi:10.6342/NTU202504505

Chapter III Methods

3-1. Data collection and outcome measurement

In the present study, all patients included in the clinical group were previously given a diagnosis based on the criteria established in the Diagnostic and Statistical Manual of Mental Disorders-V-TR (DSM-5-TR) [35-36]. For example, the DSM-5-TR defines autism spectrum disorder (ASD) as involving persistent deficits in social communication across multiple environments, as outlined in the relevant diagnostic criterion. Assessments of comorbid psychiatric diagnoses and the development of treatment plans were completed by board-certified child psychiatrists. The main caregivers of the included participants received assistance from rehabilitation therapists with gathering sociodemographic and rehabilitation-clinical information and completing several forms.

Assessments of DD symptoms were conducted by a rehabilitation physician who used the Rehabilitation Developmental Evaluation Form. In the outpatient department (OPD) of the study hospital, children with DD or child-and-adolescent psychiatry patients typically received rehabilitation therapy, and a structured data form was used to update their medical service records, which included information pertaining to the frequencies of **occupational therapy services (OTS), physical therapy services (PTS), and**

doi:10.6342/NTU202504505

speech therapy services (STS). The dataset used in this study comprised the medical records of the outpatients who visited the rehabilitation clinic of the study hospital with suspected DD between January 1, 2012, and December 31, 2016. The Institutional Review Board of En Chu Kong Hospital reviewed the above documents and approved the study on 2024/07/23 (ECK-IRB Number: ECKIRB1130501). This approval is valid until 2025/07/22. To protect patient information and confidentiality, no subject names were collected. Each patient was assigned an anonymized study ID. The International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes were used to define DD. The patients' records extracted from the dataset included age, sex, and the frequencies of OTS, PTS, and STS received. Specific DD problems and disabilities were determined using a comprehensive literature review and after a consensus was reached by rehabilitation physicians and child psychiatry specialists. The International Classification of Diseases, 10th Revision, Clinical Modification (ICD-10-CM) codes were used to identify various types of DD [37–39]. **Figure 1**. illustrates the participant selection process as a flow diagram. This study identified 2,552 outpatients under 12 years of age who made one or more OPD visits. Among these patients, 1,719 (67.4%) had DD. The total number of OPD visits was 34,862. **Table 2.** presents the demographic and clinical characteristics of the patients with DD. Because of the

doi:10.6342/NTU202504505

hospital's strict documentation flowchart, the outpatient medical records from which our dataset was derived were highly accurate, with minimal missing data and few unmeasured confounders.

The present study was approved by the ethics committee of En Chu Kong Hospital prior to data collection. Informed consent was waived by the committee because of the de-identification and non-interventional design of the present study.

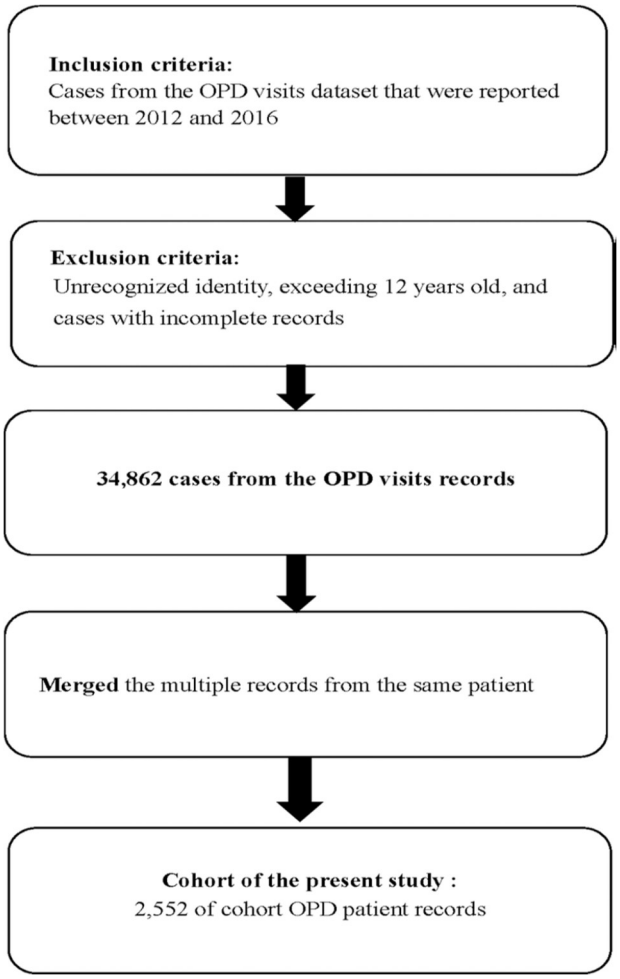


Figure 1. Flow diagram for generating the study dataset.

doi:10.6342/NTU202504505

Table 2. Demographic and clinical characteristics of patients with DD (n = 2,552).

Characteristics	Num. (%)	Num. (%)	Chi-squared test P value
Developmental Delay (DD)	**DD**	**Non-DD**	
	1,719 (67.4)	833 (32.6)	
Gender	**Men**	**Women**	**P < 0.001*** **
	1,778 (69.6)	774 (30.4)	
Age in months (Mean ± SD)	**60.5 ± 30.7**		
OTS	**Yes**	**No**	**P < 0.001*** **
	2,152 (84.3)	400 (15.7)	
PTS	**Yes**	**No**	**P < 0.001*** **
	1,624 (63.6)	928 (36.3)	
STS	**Yes**	**No**	**P < 0.001*** **
	1570 (61.5)	982 (38.5)	

NOTE. The p-values were calculated based on the $\chi 2$ test of independence. **Abbreviations**: standard deviation (SD).

3-2. Experimental procedures

The present study extracted information from OPD records, including data on demographic characteristics, such as sex and age, and the frequencies of therapy services used (OTS, PTS, and STS). The patients were divided into two groups, namely a DD group and a non-DD group. Data preprocessing included removal of incomplete records, normalization of continuous variables, and standardization of therapy frequencies to ensure cross-patient comparability. **Figure 2**. presents the experimental procedure that was employed to assess the performance of several prediction models.

doi:10.6342/NTU202504505

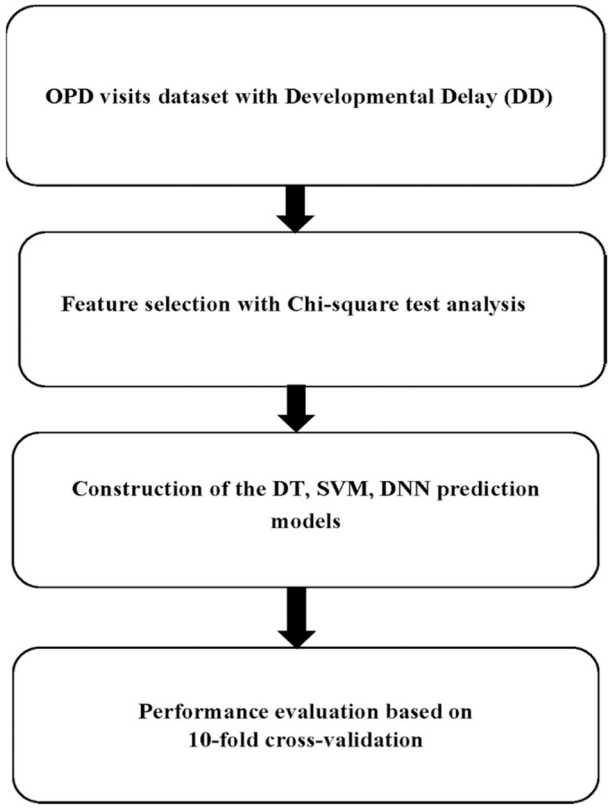


Figure 2. The experimental procedure.

3-3. Feature selection

In this study, we included four features in our dataset: sex and the frequencies of OTS, PTS, and STS. In this respect, the frequency of a particular therapy service was defined to be the average number of sessions received per patient per year. Then, we conducted chi-squared tests to determine whether a feature was correlated with the outcome variable [40–42]. For the categorical feature "sex," we carried out the chi-squared test of independence. For the frequencies of OTS, PTS, and STS, we performed chi-squared tests on discretized distributions (e.g., quartiles) as a goodness-of-fit framework, with

doi:10.6342/NTU202504505

the null hypothesis specifying that the average frequency among patients with DD equals that among patients without DD [43]. Additionally, logistic regression with forward selection was employed to confirm predictor significance. Quartiles of therapy frequency were used to estimate odds ratios (ORs). **Table 3**. shows the p-values obtained from these tests. Accordingly, we included sex and the frequencies of OTS, PTS, and STS to build the prediction models.

Table 3. Results of the feature selection by logistic regression (with odds ratios).

Variables	P value	OR (Q4 vs Q1)	95% CI
Gender	0.000022	1.84	1.45–2.33
OTS	0.000000	4.37	3.12–6.11
PTS	0.000000	3.92	2.77–5.55
STS	0.000000	4.95	3.44–6.98

3-4. Development of prediction models and performance evaluation

In this study, we investigated the prediction performance of three categories of machine learning models, namely the **decision tree (DT)** models [44–46], the **support vector machine (SVM)** models [47], and the **deep neural network (DNN)** models [48]. The DT models are preferred by many clinicians due to the explicit, human-readable decision rules output by the algorithm. On the other hand, the SVM models and the DNN models are two categories of the most advanced machine learning models that can generally outperform the DT models due to the nonlinear transformations

doi:10.6342/NTU202504505

invoked in the prediction process. However, the same nonlinear transformations also make it almost impossible for a user to comprehend how the prediction is made. As a result, many clinicians are reluctant to trust the models that work like a black box**.** Therefore, it is of interest to investigate how the performance of alternative categories of machine learning models compares. If the performance of the DT models observed in the experiments is comparable to that of the advanced machine-learning models—as observed in our recent studies [49,50]—then DT models are favored because they output explicit decision rules.

In order to obtain comprehensive pictures of how each category of prediction models performed, we employed alternative parameter settings to generate prediction models with different performance characteristics. **Table 4.** provides a summary of the software packages and alternative parameter settings employed to build the prediction models. Then, we conducted 10-fold cross-validation to evaluate the performance characteristics of each prediction model generated [51–53]. The performance metrics considered in this study include accuracy, sensitivity, specificity, positive predictive value (PPV; also known as precision), and F1 score**.** The F1 score, which is the harmonic mean of the sensitivity and the PPV, is commonly employed in machine-learning research and has increasingly been employed in biomedical research [54]. Furthermore, for each category

doi:10.6342/NTU202504505

of prediction models (e.g., DT, SVM, or DNN), we evaluated overall performance based on the area under the receiver operating characteristic (ROC) curve [55-56]**.** To generate each ROC curve, we selected, at every sensitivity level, the configuration within that model family that delivered the highest F1 score.

Table 4. Software packages and parameter settings employed to build the models.

Model	Programming Language	Package	Parameters
DT	Python	pandas sklearn	Split = "information" Prior = 0.01 ~ 0.9 with 0.008 step size, cp = [0.05,0.04,0.03,0.02,0.01]
DNN	Python	Tensor Flow	Input neurons = [4] for 4 features set, Hidden neurons = [5,10,27] Hidden layer = [3,5,9]
SVM	Python	sklearn svm	Kernel = ['linear','rbf','poly'], C= [0.01,0.1,1,3,10,15,20]; gamma = 'auto' Probability = True; Best for rbf-kernal (cost = 10, gamma = 0.25, epsilon = 0.1)

doi:10.6342/NTU202504505

Chapter IV Results

The dataset comprised 2,552 children and 34,862 visits, with a mean age of 72.34 months. Gender distribution showed a higher proportion of males, consistent with prior prevalence studies [57]. Children with DD demonstrated significantly greater utilization of OTS, PTS, and STS compared with non-DD peers ($p < 0.001$) shown in **Table 2.** Weighted modeling and stratified sampling approaches confirmed that this gender imbalance did not undermine classification reliability. Sensitivity analyses excluding age revealed minimal reduction in model accuracy, confirming therapy frequency as the dominant predictor.

In **Table 3**., odds Ratios showed children with high therapy frequencies had over fourfold increased risk of DD.

Figure 4. shows ROC curves and the corresponding areas under the curves (AUCs) of the DT, SVM, and DNN models. **Table 5.** shows the detailed performance data of the models that delivered sensitivities at the 0.80 level and at the 0.90 level. It is observed that the DNN models and the DT models outperformed the SVM models in terms of AUC. On the other hand, as shown in **Table 5.**, the performance comparison showed DT sensitivity = 0.902, PPV = 0.723, and F1 = 0.803. DNN achieved higher AUC but lower PPV, while SVM underperformed across metrics. If a high level of sensitivity is

doi:10.6342/NTU202504505

desirable, then the DT models significantly outperformed the DNN models and the SVM models in terms of the F1 score, which is the harmonic mean of the sensitivity (also called recall) and the positive predictive value (PPV, also called precision). Based on the data shown in **Table 5.**, it is conceivable that the DT model that delivered the sensitivity at the 0.90 level is the favorite choice due to two reasons. Firstly, the PPV with this particular DT model is significantly higher than the PPVs with the SVM model and the DNN model that delivered the same level of sensitivity. Therefore, in clinical applications, the number of false positive predicted by this DT model should be significantly lower than the numbers of false positive predicted by the SVM model and the DNN model with the same level of sensitivity. Secondly, the PPV with this DT model is almost the same as the PPV with the DT model that delivered the sensitivity at the 0.80 level. Accordingly, in the subsequent discussions, we will focus on the DT model that delivered the sensitivity at the 0.90 level.

Figure 3. shows the structure of the DT model generated by feeding our dataset into the software package and with cp and prior set to 0.01 and 0.55 respectively. According to our performance evaluation, this DT model should be able to deliver a sensitivity at the 0.90 level and a PPV above 0.70. The top-down path, following the red arrows, illustrates how the prediction for a subject with sex = female, f_OTS = 60, f_PTS = 30,

doi:10.6342/NTU202504505

and f_STS = 50 is made. The prediction made is positive, i.e., the subject suffers from DD, as the path ends at a red node. On the other hand, a subject is predicted to be negative, if the path corresponding to the subject's feature values ends at a blue node. The “n+” and “n-“ values associated with each node respectively specify the percentages of positive subjects and negative subjects among all the subjects that meet the criteria specified along the path to the node. In fact, a user can figure out the probability that a subject is positive or negative by examining the n+ and n- of the leaf node that the feature values of this subject fit into.

These results confirm therapy frequencies are strong, low-cost predictors, and that DT combines accuracy with interpretability, making it suitable for clinical adoption.

doi:10.6342/NTU202504505

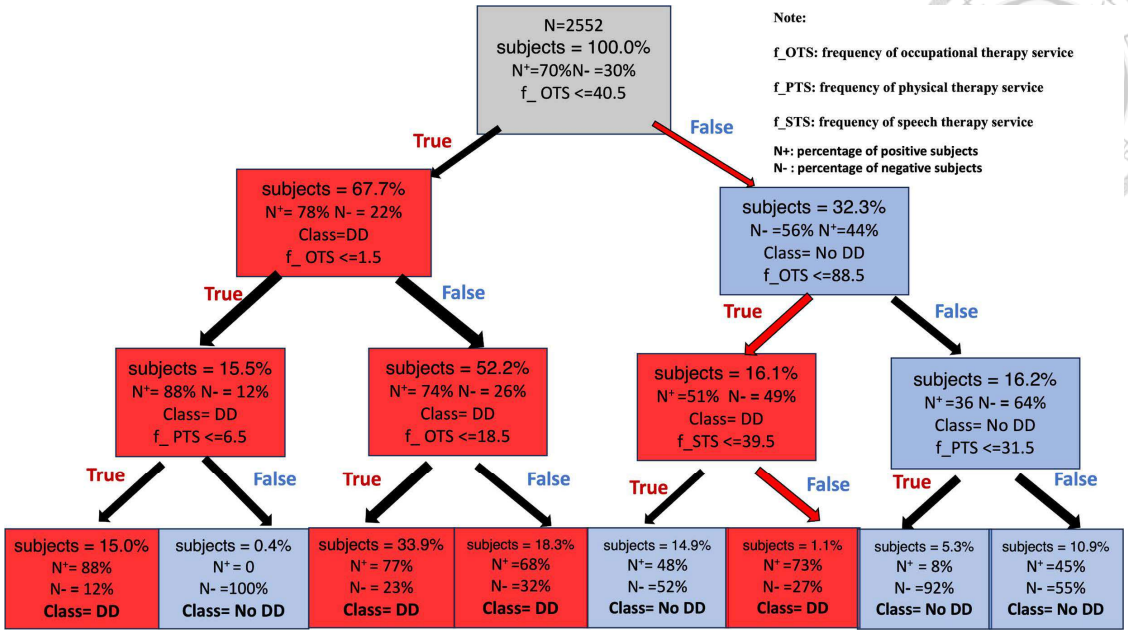


Figure 3. The structure of the DT model generated by feeding our dataset into the software package and with cp and prior set to 0.01 and 0.55, respectively.

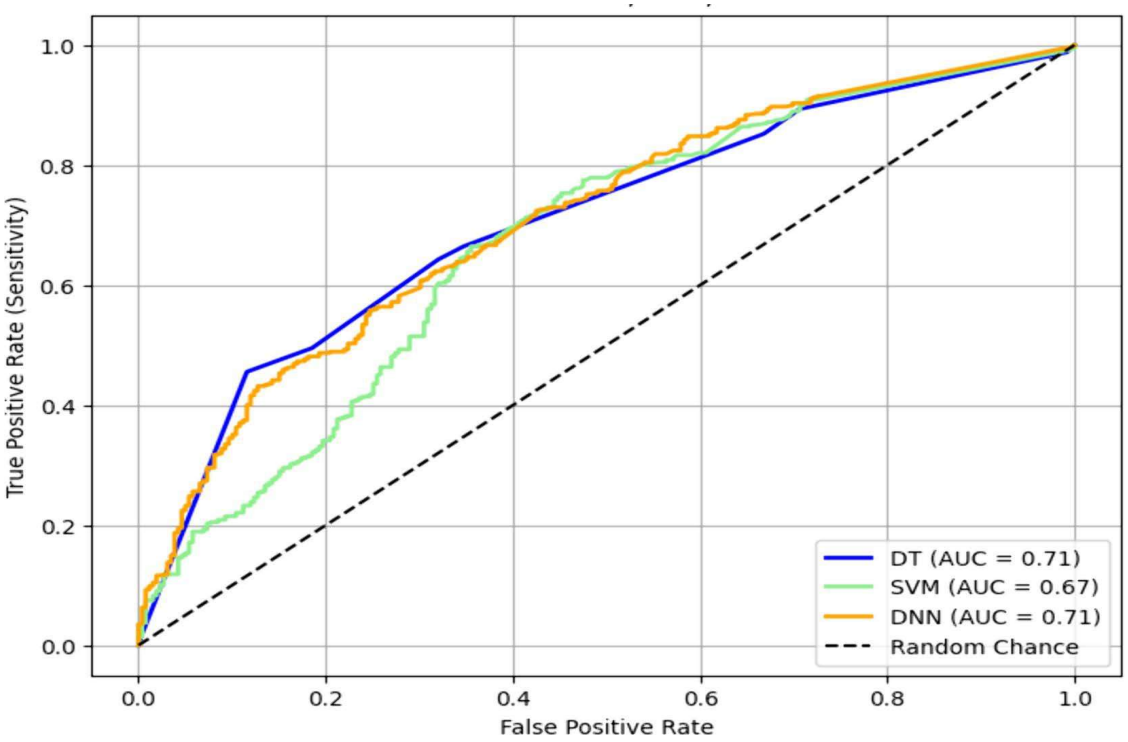


Figure 4. ROC curves of the DNN, DT, SVM models.

doi:10.6342/NTU202504505

Table 5. Detailed performance characteristics of alternative prediction models.

Target Sensitivity	Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score
0.8	DT	0.650	0.802	0.337	0.734	0.744
	SVM	0.597	0.802	0.488	0.454	0.580
	DNN	0.663	0.808	0.587	0.509	0.624
0.9	**DT**	**0.701**	**0.902**	**0.289**	**0.723**	**0.803**
	SVM	0.487	0.898	0.269	0.395	0.548
	DNN	0.616	0.904	0.464	0.472	0.620

doi:10.6342/NTU202504505

Chapter V Discussion

The findings highlight therapy frequencies as clinically valuable predictors of DD. Unlike high-cost imaging or genetic testing, therapy records are universally available in outpatient care and can support cost-effective screening. The DT model's explicit rules enhance clinical trust and usability. This transparency addresses the black-box limitations of models such as DNN, providing interpretable decision pathways that clinicians can readily apply.

Model evaluation extended beyond conventional accuracy and AUC. The Precision-Recall curves [58] provided insight under imbalance, while Odds Ratios linked predictions to clinically interpretable risk. Confusion matrices further demonstrated DT's superior balance of true and false classifications.

Bias due to male predominance was addressed by weighted modeling, confirming DT stability. Excluding age did not compromise performance, underscoring therapy frequency as the dominant predictor. Future incorporation of high-dimensional features may improve precision but could reduce interpretability.

In this study, we have investigated how the frequencies of therapies can be exploited to build machine learning based prediction models for identifying children with development delay. Based on the experimental results observed, it is conceivable that

doi:10.6342/NTU202504505

the proposed approach can be widely exploited in clinical practices due to several reasons. Firstly, the performance observed with the prediction models developed in this study should meet the criteria acceptable by most physicians. For example, based on our experimental results, we can anticipate that the DT model shown in **Table 5.** can identify about 90.0% of the subjects who will develop DD in the future, while about 72% of the subjects predicted to be positive are actually true positives. Secondly, the features employed to build the prediction models can be obtained with essentially no costs. Therefore, the prediction models can be exploited to screen the subjects who may develop DD before advanced and costly diagnoses are carried out.

The experimental results also demonstrate that for the applications targeted by this study we do not need to trade performance for the interpretability of the prediction model. The F1 scores presented in **Table 5.** show that the DT models that delivered the sensitivity at the 0.90 level and at the 0.80 level outperformed the DNN models and the SVM model that delivered the sensitivity at the same level. For most applications, it is typical that advanced machine learning based prediction models such as the DNN models and the SVM models outperform the DT models due to the non-linear transformations invoked. However, the non-linear transformations invoked also make it almost impossible for a user to figure out how the prediction is made. Fortunately, for

doi:10.6342/NTU202504505

our applications, we do not need to trade performance for the interpretability of the prediction model.

The DT structure shown in **Figure 3** illustrates how a user can examine the structure to figure out the decision rules followed by the prediction model to make predictions. Furthermore, the ratio of between the number of positive subjects and the number of negative subjects at each leaf node specifies how likely a subject that meets the criteria corresponding to the path to this particular leaf node develops DD. For example, the probability that the subject with sex = female, f_OTS = 60, f_PTS = 30, and f_STS = 50 develops DD is 0.73. In clinical practice, a physician can refer to this specific probability and his/her clinical experiences to make the final diagnosis.

In summary, the major finding due to this study is that the frequencies of the therapies that a child has received provide valuable information for predicting whether the child suffers from DD. Due to the performance observed in the experiments and the fact that these features can be obtained essentially without any cost, it is conceivable that the prediction models built accordingly can be wide exploited in clinical practices and significantly improve the treatment outcomes of the children who develop DD. Though the study was based on a dataset collected in a hospital in Taiwan, we anticipate that the

doi:10.6342/NTU202504505

proposed method can be exploited to build accurate prediction models for populations in different countries with various race groups.

Limitations

Several limitations of this study should be noted. Firstly, this retrospective study relies on data extracted from the outpatient (OPD) database with children under 12 years old. Consequently, the findings may not be generalized for the other age groups. Secondly, the prediction models developed were solely based on the data collected from a hospital in Taiwan and its applicability to other hospitals has not been validated. Thirdly, the dataset employed in this study was derived from the clinical records in the OPD and therefore these patients were likely to already have DD conditions. Fourth, the restricted feature set limited granularity. Only therapy

frequencies, sex, and age were included; important factors such as comorbidities, socioeconomic variables, family history, and longitudinal clinical data were unavailable. Fifth, despite the DT achieving a sensitivity of approximately 90%, around 10% of DD cases were misclassified as false negatives. This limitation is clinically significant, as it could delay recognition and treatment for a subset of children. Finally, it is observed that there were significantly more male patients than the female patients, which

doi:10.6342/NTU202504505

conforms with previous findings [59,60]. Therefore, stratified sampling based on gender was not carried out.

Limitations include reliance on a single hospital dataset and retrospective design. Future validation across hospitals and countries is necessary to enhance generalizability [61,62]. The use of ICD-9-CM and ICD-10-CM coding ensures compatibility, but variability in clinical documentation must still be considered.

Ethical considerations include safeguarding patient privacy, ensuring transparent predictive processes, and establishing follow-up protocols for the ~10% of cases potentially missed by the model. Multi-stakeholder collaboration among clinicians, data scientists, and policymakers will be vital for responsible integration.

doi:10.6342/NTU202504505

Chapter VI Conclusion

This study demonstrates that therapy utilization frequencies—specifically occupational, physical, and speech therapies—are powerful, low-cost predictors of DD. The findings confirm that routinely collected rehabilitation data can serve as pragmatic indicators for scalable early screening, offering a cost-effective alternative to resource-intensive diagnostic modalities.

Among the models evaluated, the DT consistently achieved a clinically meaningful balance of sensitivity and positive predictive value, while also providing interpretability that is essential for clinical adoption. The transparent decision rules embedded in DT structures allow physicians to integrate computational outputs with their clinical expertise, thereby supporting diagnostic reasoning and therapeutic planning. Compared with more complex models such as the SVM and DNN, the DT demonstrated superior clinical usability despite similar or slightly lower discriminative performance.

By showing that low-cost, readily available clinical features can be effectively translated into interpretable ML models, this dissertation establishes a foundation for scalable ML-based DD screening systems. Such systems have the potential to complement, rather than replace, clinical expertise, enabling earlier detection, reducing

doi:10.6342/NTU202504505

diagnostic delays, and ultimately improving developmental outcomes in pediatric populations.

doi:10.6342/NTU202504505

Chapter VII Future works

Future research should pursue both methodological and clinical advancements to enhance the robustness, interpretability, and applicability of ML-based DD screening.

1. Methodological Enhancements

Ensemble approaches such as Extreme Gradient Boosting (XGBoost) [63,64] and Random Forests should be investigated, given their robustness against noisy data and capacity to capture complex non-linear interactions. Standardized benchmarking frameworks are also necessary, comparing ML algorithms not only against one another but also against established diagnostic standards such as DSM-5–based developmental assessments. Such comparisons would provide stronger evidence of the incremental value of ML approaches over current clinical practice.

2. Expansion of Predictive Features

While this study validated therapy frequencies, age, and sex as effective predictors, expanding the feature set could substantially improve predictive power. Potential additions include comorbidity profiles, hospitalization and medication history, socioeconomic variables, and longitudinal developmental records. Incorporating ICD-9-CM and ICD-10-CM subcategories (e.g., F80–F89 neurodevelopmental disorders, G80

doi:10.6342/NTU202504505

cerebral palsy, R62 developmental delay) would enable subtype-level classification. Such integration also improves interoperability with electronic health record systems, facilitating adoption in clinical workflows.

3. Multimodal Predictive Frameworks

Future studies should explore hybrid models that combine routine clinical metadata with high-dimensional modalities such as EEG, neuroimaging, and genetic features. These multimodal approaches could balance feasibility with diagnostic precision, particularly in tertiary care or research settings where advanced diagnostic resources are available.

4. Evaluation Strategies

Performance evaluation should move beyond ROC curves alone. Precision–Recall (PR) curves offer superior insight into imbalanced datasets, while odds ratios contextualize model outputs in epidemiological terms familiar to clinicians. Establishing standardized evaluation metrics will improve methodological rigor and clinical interpretability across future studies.

5. Addressing False Negatives

Although the DT model achieved ~90% sensitivity, approximately one in ten cases remained undetected. This limitation could be mitigated through ensemble modeling,

doi:10.6342/NTU202504505

richer feature integration, and structured re-screening protocols. For example, children initially classified as non-DD but who continue to receive high-frequency therapy could be flagged for follow-up evaluation. Such safeguards are critical to reducing missed diagnoses and preventing delays in intervention.

6. Age-Specific Therapy Patterns

A particularly valuable research direction is the investigation of associations between therapy frequency and age. Identifying whether therapy utilization peaks within specific developmental stages may provide empirical support for the principle of early intervention. Furthermore, analyzing the age at diagnosis, therapy types received, and the period of highest therapy concentration could refine understanding of the critical timeframe for intervention.

7. Validation and Generalizability

Prospective, multicenter, and cross-national validation is essential to confirm generalizability. Differences in healthcare systems, cultural contexts, and access to therapy must be considered. Cross-institutional collaborations will help prevent overfitting to local patterns and ensure equitable applicability of ML-based DD screening tools worldwide.

doi:10.6342/NTU202504505

8. Ethical and Clinical Integration

Future implementations must prioritize ethical safeguards. Clinical integration should include structured protocols for follow-up of false negatives, safeguards for patient privacy, and mechanisms to monitor model drift as therapy practices evolve. Advances in explainable AI frameworks may further refine interpretability, ensuring that predictive systems complement—rather than replace—clinical expertise.

Summary

By advancing methodological rigor, expanding feature sets, integrating multimodal data, and embedding ethical safeguards, future research can establish ML-based DD screening as both a responsible and transformative tool in pediatric healthcare.

doi:10.6342/NTU202504505

References

1. Choi WW, McBride CA, Bourke C, Borzi P, Choo K, Walker R, et al. Long-term review of sutureless ward reduction in neonates with gastroschisis in the neonatal unit. J Pediatr Surg. 2012;47(8):1516–20. https://doi.org/10.1016/j.jpedsurg.2012.01.010 PMID: 22901910

2. Shevell M, Ashwal S, Donley D, Flint J, Gingold M, Hirtz D, et al. Evaluation of the child with global developmental delay and intellectual disability: recommendations of the Canadian Paediatric Society and the American Academy of Neurology. Paediatr Child Health. 2011;16(7):422-8.

3. Noritz GH, Murphy NA. Motor delays: Early identification and evaluation. Pediatrics. 2013;131(6): e2016–27.

4. Park ES, Kim DY, Kim AR, Rha DW, Park CI. Clinical characteristics and rehabilitation potential in children with cerebral palsy based on MRI classification system. Ann Rehabil Med. 2024;48(2):112-22.

5. Whyte J, Dijkers MP, Hart T, Zanca JM, Packel A, Ferraro M, et al. Treatment taxonomy for rehabilitation: past, present, and prospects. Arch Phys Med Rehabil. 2014;95(1 Suppl): S6-16.

doi:10.6342/NTU202504505

6. Rosenbaum P, Paneth N, Leviton A, Goldstein M, Bax M, Damiano D, et al. A report: the definition and classification of cerebral palsy April 2006. Dev Med Child Neurol Suppl. 2007; 109:8-14.

7. Katuwal G. Machine learning-based autism detection using brain imaging. Rochester (NY): Rochester Institute of Technology. 2017. https://scholarworks.rit.edu

8. Smythe T, Zuurmond M, Tann CJ, Gladstone M, Kuper H. Early intervention for children with developmental disabilities in low and middle-income countries - the case for action. Int Health. 2021;13(3):222–31. https://doi.org/10.1093/inthealth/ihaa044 PMID: 32780826

9. Reynolds AJ, Temple JA, Robertson DL, Mann EA. Long-term effects of an early childhood intervention on educational attainment and juvenile arrest: a 15-year follow-up of low-income children. *JAMA*. 2001;285(18):2339–46. https://doi.org/10.1001/jama.285.18.2339 PMID: 11343481

10. Estes A, Munson J, Rogers SJ, Greenson J, Winter J, Dawson G. Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry*. 2015;54(7):580–7. https://doi.org/10.1016/j.jaac.2015.04.005 PMID: 26088663

doi:10.6342/NTU202504505

11. Fuller EA, Kaiser AP. The effects of early intervention on social communication outcomes for children with autism spectrum disorder: a meta-analysis. *J Autism Dev Disord*. 2019;49(3): 1257–75. https://doi.org/10.1007/s10803-018-3844-7 PMID: 30632057

12. Hirve Y, Bhatia R, Richter LM, et al. Effect of early childhood development interventions delivered through health services on cognitive outcomes: pooled analysis of trials. *Arch Dis Child*. 2023;108(4):247–54. https://doi.org/10.1136/archdischild-2022-324844 PMID: 36848782

13. Reichow B, Hume K, Barton EE, Boyd BA. Early intensive behavioral intervention (EIBI) for young children with autism spectrum disorders. *Cochrane Database Syst Rev*. 2018;(5):CD009260. https://doi.org/10.1002/14651858.CD009260.pub3 PMID: 29770488

14. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press; 2014.

15. Hair FJ, Sarstedt M. Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing. J Mark Theory Pract. 2021;29(1):65–77.

doi:10.6342/NTU202504505

16. Wall DP, Kosmicki J, Deluca TF, Harstad E, Fusaro VA. Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl Psychiatry. 2012;2(4): e100. https://doi.org/10.1038/tp.2012.10 PMID: 22832900

17. Mueller A, Candrian G, Kropotov J, Ponomarev V, Baschera G. Classification of ADHD patients on the basis of independent ERP components using a machine learning system. Nonlinear Biomed Phys. 2010; 4:1.

18. Bedi G, Carrillo F, Cecchi GA, Slezak DF, Sigman M, Mota NB, et al. Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. 2015; 1:15030. https://doi.org/10.1038/npjschz.2015.30 PMID: 27336038

19. Bishop-Fitzpatrick L, Movaghar A, Greenberg JS, Page D, DaWalt LS, Brilliant MH, et al. Using machine learning to identify patterns of lifetime health problems in decedents with autism spectrum disorder. Autism Res. 2018; 11(8):1120-8. https://doi.org/10.1002/aur.1960 PMID: 29734508

20. Herring S, Gray K, Taffe J, Tonge B, Sweeney D, Einfeld S. Behaviour and emotional problems in toddlers with pervasive developmental disorders and developmental delay: associations with parental mental health and family functioning. J Intellect Disabil Res. 2006;50(Pt 12): 874-82. https://doi.org/10.1111/j.1365-2788.2006.00904.x PMID: 17100948

doi:10.6342/NTU202504505

21. Altay O, Ulas M. Prediction of the autism spectrum disorder diagnosis with linear discriminant analysis classifier and K-nearest neighbor in children. In: 2018 6th International Symposium on Digital Forensic and Security (ISDFS). IEEE. 2018.

22. Büyükoflaz F, Öztürk A. Early autism diagnosis of children with machine learning algorithms. In: 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE. 2018.

23. Dvornek NC, Ventola P, Duncan JS. Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018; Washington (DC), USA: IEEE; 2018.8

24. Liao D, Lu H. Classify autism and control based on deep learning and community structure on resting-state fMRI. In: 2018 IEEE 10th International Conference on Advanced Computational Intelligence (ICACI); 2018; Xiamen, China: IEEE; 2018.

25. Dekhil O, Ismail M, Shalaby A, et al. A novel CAD system for autism diagnosis using structural and functional MRI. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017); 2017; Melbourne (VIC), Australia: IEEE; 2017.

doi:10.6342/NTU202504505

26. Jiao Y, Lu Z. Predictive models for ASD based on multiple cortical features. In: 2011 IEEE 8th International Conference on Fuzzy Systems and Knowledge Discovery; 2011; Shanghai, China: IEEE; 2011. p.1611-5.

27. Heinsfeld AS, Franco AR, Craddock RC, Buchweitz A, Meneguzzi F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. Neuroimage Clin. 2017; 17:16-23. https://doi.org/10.1016/j.nicl.2017.08.017 PMID: 29034163

28. Bone D, Bishop SL, Black MP, Goodwin MS, Lord C, Narayanan SS. Use of machine learning to improve autism screening and diagnostic instruments: effectiveness, efficiency, and multi-instrument fusion. J Child Psychol Psychiatry. 2016;57(8): 927-37. https://doi.org/10.1111/jcpp.12559 PMID: 27090613

29. Jin Y, Wee C-Y, Shi F, Thung K-H, Ni D, Yap P-T, et al. Identification of infants at high-risk for autism spectrum disorder using multiparameter multiscale white matter connectivity networks. Hum Brain Mapp. 2015;36(12):4880–96. https://doi.org/10.1002/hbm.22957 PMID: 26368659

30. Kim B-M, Kang B-Y, Kim H-G, Baek S-H. Prognosis prediction for Class III malocclusion treatment by feature wrapping method. Angle Orthod. 2009;79(4):683. https://doi.org/10.2319/071508-371.1 PMID: 19537866

doi:10.6342/NTU202504505

31. Song C, Jiang Z-Q, Liu D, Wu L-L. Application and research progress of machine learning in the diagnosis and treatment of neurodevelopmental disorders in children. Front Psychiatry. 2022; 13:960672. https://doi.org/10.3389/fpsyt.2022.960672 PMID: 36090350

32. Megerian JT, Dey S, Melmed RD, Coury DL, Lerner M, Nicholls CJ, et al. Evaluation of an artificial intelligence-based medical device for diagnosis of autism spectrum disorder. NPJ Digit Med. 2022;5(1):57. https://doi.org/10.1038/s41746-022-00598-6 PMID: 35513550

33. Maharjan J, Garikipati A, Dinenno FA, Ciobanu M, Barnes G, Browning E, et al. Machine learning determination of applied behavioral analysis treatment plan type. Brain Inform. 2023;10(1):7. https://doi.org/10.1186/s40708-023-00186-8 PMID: 36862316.

34. Tariq Q, Fleming SL, Schwartz JN, Dunlap K, Corbin C, Washington P, et al. Detecting Developmental Delay and Autism Through Machine Learning Models Using Home Videos of Bangladeshi Children: Development and Validation Study. J Med Internet Res. 2019;21(4): e13822. https://doi.org/10.2196/13822 PMID: 31017583

doi:10.6342/NTU202504505

35. American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5th ed. ed. Washington (DC): American Psychiatric Association. 2013.

36. El-Baz F, El-Aal MA, Kamal TM, Sadek AA, Othman AA. Study of the C677T and 1298AC polymorphic genotypes of MTHFR Gene in autism spectrum disorder. Electron Physician. 2017;9(9):5287–93. https://doi.org/10.19082/5287 PMID: 29038711

37. World Health Organization. International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). Geneva: WHO; 1979.

38. World Health Organization. International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). Geneva: WHO; 1990.

39. Centers for Disease Control and Prevention. ICD-10-CM official guidelines for coding and reporting. Atlanta (GA): CDC; 2016.

40. Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med. 2019; 112:103375. https://doi.org/10.1016/j.compbiomed.2019.103375 PMID: 31382212.

41. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer. 2013.

doi:10.6342/NTU202504505

42. Cho S, Hong H, Ha B. A hybrid approach based on the combination of variable selection using decision trees and case-based reasoning using the Mahalanobis distance for bankruptcy prediction. Expert Syst Appl. 2010;37(5): 3482-8.

43. Kallenberg W, Oosterhoff J, Schriever B. The number of classes in chi-squared goodness-of-fit tests. J Am Stat Assoc. 1985;80(392): 959-68.

44. Maimon O, Rokach L. Data mining with decision trees: theory and applications. 2nd ed. ed. Hackensack (NJ): World Scientific. 2014.

45. Bai J, Li Y, Li J, Jiang Y, Xia S. Rectified decision trees: towards interpretability, compression, and empirical soundness. arXiv. https://arxiv.org/abs/1903.05965. 2019. Accessed 2025 March 17.

46. Tharwat A. Classification assessment methods. Appl Comput Inform. 2020;17(2): 168-92.

47. Xue, Hui, Qiang Yang, and Songcan Chen. "SVM: Support vector machines." *The top ten algorithms in data mining*. Chapman and Hall/CRC, 2009. 51-74.

48. Desai VS, Crook JN, Overstreet GAJ. A comparison of neural networks and linear scoring models in the credit union environment. Eur J Oper Res. 1996;95(1): 24-37.

49. Ho T-S, Weng T-C, Wang J-D, Han H-C, Cheng H-C, Yang C-C, et al. Comparing machine learning with case-control models to identify confirmed dengue cases. PLoS

doi:10.6342/NTU202504505

Negl Trop Dis. 2020;14(11):e0008843. https://doi.org/10.1371/journal.pntd.0008843 PMID: 33170848

50. Chiu H-YR, Hwang C-K, Chen S-Y, Shih F-Y, Han H-C, King C-C, et al. Machine learning for emerging infectious disease field responses. Sci Rep. 2022;12(1):328. https://doi.org/10.1038/s41598-021-03687-w PMID: 35013370

51. Fushiki T. Estimation of prediction error by using K-fold cross-validation. Stat Comput. 2011;21(2):137-46.

52. Wong TT, Yeh PY. Reliable accuracy estimates from k-fold cross-validation. IEEE Trans Knowl Data Eng. 2019;32(8): 1586-94.

53. Munsch N, Martin A, Gruarin S, Nateqi J, Abdarahmane I, Weingartner-Ortner R, et al. Diagnostic Accuracy of Web-Based COVID-19 Symptom Checkers: Comparison Study. J Med Internet Res. 2020; 22(10): e21299. https://doi.org/10.2196/21299 PMID: 33001828

54. Lundberg SM, Lee S. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017: 30: 4765-74.

55. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143(1): 29-36. https://doi.org/10.1148/radiology.143.1.7063747 PMID: 7063747

doi:10.6342/NTU202504505

56. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006; 27(8): 861-74.

57. Herring S, Gray K, Taffe J, Tonge B, Sweeney D, Einfeld S. Behaviour and emotional problems in toddlers with pervasive developmental disorders and developmental delay: associations with parental mental health and family functioning. J Intellect Disabil Res. 2006;50(Pt 12): 874-82. https://doi.org/10.1111/j.1365-2788.2006.00904.x PMID: 17100948

58. Saito T, Rehmsmeier M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One.

59. Russell G, Steer C, Golding J. Social and demographic factors that influence the diagnosis of autistic spectrum disorders. Soc Psychiatry Psychiatr Epidemiol. 2011;46(12):1283-93. https://doi.org/10.1007/s00127-010-0294-z PMID: 20938640

60. Fombonne E. Epidemiology of pervasive developmental disorders. Pediatr Res. 2009; 65(6): 591-8. https://doi.org/10.1203/PDR.0b013e31819e7203 PMID: 19218885. customer carecustomer careauthor billing.

61. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.

62. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

doi:10.6342/NTU202504505

63. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017; 30:3149–5.

64. Zhou Z-H. Ensemble methods: foundations and algorithms. Boca Raton (FL): Chapman & Hall/CRC; 2012.

doi:10.6342/NTU202504505