Final Project Research Report

Investigating the Relationship Between Risk Factors and Neurodevelopmental Disabilities in Children Using Machine Learning.

Abstract:

This study investigates the relationship between participant characteristics (risk factors) and neurodevelopmental disabilities (NDDs) in children aged 3–17 years using data from the National Survey of Children's Health (NSCH). The dataset was analyzed to explore associations between key variables, including demographic, biological, and environmental factors, with NDD outcomes such as ADHD, behavioral problems, and learning disabilities.

Machine learning techniques, specifically Random Forest models, were utilized to forecast NDD results and evaluate the significance of the variables. In contrast, statistical techniques, such as Chi-Square tests, were utilized to investigate correlations. The Random Forest model showed excellent prediction ability with an accuracy of 95% and an AUC score of 0.976. In addition to environmental factors, including family support and neighborhood safety, age, screen use, and physical activity levels were significant predictors. These results imply that contextual and individual factors significantly influence neurodevelopmental outcomes.

This work emphasizes the value of early identification and intervention to reduce the risks of NDD and shows how machine learning techniques may be used to identify essential risk variables. This work offers insights that can guide public health initiatives and interventions to enhance neurodevelopmental outcomes by integrating variable relevance assessments with predictive modeling. To improve risk prediction and provide specialized preventative strategies, future studies should investigate the interaction of genetic and environmental factors in more detail.

Background:

Neurodevelopmental disorders (NDD), or neurodevelopmental delays, involve delayed skill development in infants and children. These disorders result from factors that affect the nervous system. Proper central nervous system (CNS) functioning is crucial in child development. Various biological and environmental factors can influence the CNS, potentially leading to neurodevelopmental disorders(1). These disorders may arise from genetic abnormalities, chromosomal defects, infections, perinatal brain injuries, or disruptions in neuronal migration, all of which can impair brain development. Environmental factors like poverty, malnutrition, and inadequate cognitive stimulation are significant contributors to atypical neurodevelopment(2). Biological and environmental factors often interact with environmental risks, exacerbating biological vulnerabilities. In contrast, positive influences like adequate nutrition, stimulating environments, and maternal factors like education level and responsiveness can support healthy development and improve outcomes despite biological challenges(3). With technological advancements, particularly in machine learning (ML), innovative approaches are being developed to improve the accuracy and efficiency of diagnosing NDDs. ML algorithms can analyze vast amounts of complex data, uncovering patterns and relationships that may be challenging for humans to detect. This enables the creation of more precise and efficient diagnostic tools to aid clinicians in early detection, which is crucial during critical developmental periods when timely interventions can greatly enhance long-term outcomes. Numerous studies have applied various ML techniques to analyze different data types for diagnosing and predicting NDDs, prioritizing accuracy and cost-effectiveness. These approaches primarily utilize supervised learning methods such as regression, support vector machines (SVMs), decision trees, artificial neural networks (ANNs), Bayesian logic, and unsupervised techniques like clustering, association rules, and dimensionality reduction. While less common, semi-supervised and reinforcement learning methods also hold potential in this domain(4).

This study analyzes associations between significant risk factors and various NDD outcomes using statistical and machine-learning approaches. The primary objectives are to identify predictors of NDDs, evaluate their relative importance, and assess the accuracy of ML-based models in predicting NDD outcomes. Specifically, we apply Random Forest—a robust and interpretable machine learning algorithm—to predict NDD outcomes and determine the most influential variables. This analysis aims to provide insights into modifiable risk factors, enabling early detection and intervention strategies to improve outcomes in affected children. By combining statistical analysis with machine learning techniques, this study contributes to the growing body of research aimed at understanding and predicting NDDs using large-scale datasets. The findings are expected to offer valuable guidance for clinicians, public health professionals, and policymakers in identifying risk factors and designing preventive strategies to mitigate the impact of neurodevelopmental disorders.

Study Design

Aim:

The primary aim of this study is to analyze the associations between participant characteristics and neurodevelopmental disorders (NDD) outcomes using statistical and machine learning techniques. Specifically, this study seeks to:

1. Identify relevant risk factors associated with NDD outcomes by selecting at least 20 independent variables. These include individual-level factors (e.g., age, gender, race/ethnicity, BMI, physical activity, and screen time), clinical conditions (e.g., comorbidities), and family environment factors (e.g., smoking exposure, neighborhood safety, and emotional support).
2. Using machine learning methods, Investigate the relationships between these risk factors and NDD outcomes.
3. Highlight variable importance and identify key predictors influencing NDD outcomes to inform early detection and intervention strategies.

Population Selection:

The study utilized data from the National Survey of Children's Health (NSCH) 2022, comprising 54,103 participants aged 3–17. Relevant variables were identified based on keywords associated with neurodevelopmental disorders (NDDs) and participant characteristics, including demographics, clinical conditions, and family environment factors, ensuring a comprehensive analysis.

Data and Materials:

The study utilized two CSV files extracted from the *NSCH 2022 dataset* provided in a zip file: one containing raw survey data (NSCH_2022e_Topical_CSV_CAHMI_DRCv2.csv) and the other containing variable descriptions (NSCH_2022e_CAHMI_DRCv2_CSV_Variable labels.xlsx). The dataset comprises 54,103 participants aged 3–17 years and includes comprehensive information on child health, demographics, and environmental factors. For the analysis, two variable subsets were used. Cerebral palsy, Down syndrome, epilepsy, Tourette syndrome, anxiety, depression, behavioral problems, developmental delays, intellectual disability, speech disorders, learning disabilities, autism spectrum disorder, and ADHD were among the 13 variables that represented neurodevelopmental disorders (NDDs) in the NDD outcomes subset.

Age, gender, race/ethnicity, BMI, physical activity, screen time, sleep hours, allergies, diabetes, heart conditions, memory problems, brain injuries, severity of blood disorders, bullying, genetic or inherited conditions, smoking environments, neighborhood safety, family emotional support, meals together, and housing conditions were among the 20 variables that may influence NDDs that were included in the risk factors subset.

Outliers, missing values, and category encoding were all addressed during data preprocessing. These resources were the foundation for investigating relationships between risk factors and NDD outcomes and using machine learning models to investigate the significance of various variables and predictive relationships. All analyses were conducted using R programming.

Statistical Analysis

Chi-square tests were conducted to identify significant relationships between neurodevelopmental disorder (NDD) outcomes and selected risk factors. The analysis used two datasets: ndd_subset1, containing 13 NDD-related variables (e.g., autism, ADHD, speech disorders), and risk_factor_subset1, comprising 20 risk factors (e.g., age, BMI, physical activity, and housing conditions).

A correlation matrix was initially computed to identify interdependencies between variables. The correlation analysis highlighted moderate to strong associations among several NDD outcomes, particularly between Tourette syndrome, anxiety, depression, and behavioral issues. For instance, Tourette syndrome was strongly correlated with anxiety (r=0.96r = 0.96r=0.96), depression (r=0.97r = 0.97r=0.97), and behavioral issues (r=0.97r = 0.97r=0.97). Similarly, risk factors like sleep hours (r=0.55r = 0.55r=0.55) and memory difficulty (r=0.55r = 0.55r=0.55) exhibited noteworthy relationships with NDD variables.

Chi-square tests confirmed statistically significant associations between multiple NDD outcomes and risk factors ($p < 0.05$). For instance:

- Cerebral palsy was significantly associated with factors like age (p=0.0009), physical activity (p=0.0005), screen time (p=0.0009), and housing (p=0.0029).
- Tourette syndrome exhibited significant associations with age (p=0.0005), sex (p=0.0005), BMI (p=0.0005), and emotional support (p=0.0005).
- Behavioral issues were linked with nearly all tested risk factors, including age (p=0.0005), race (p=0.0005), and physical activity (p=0.0005).

These analyses identified risk factors most strongly associated with specific NDD outcomes, providing insights for further predictive modeling. Statistical tests confirmed meaningful relationships, underscoring the utility of machine learning for deeper exploration of these interactions.

Machine Learning Analysis

To investigate the relationship between neurodevelopmental disorder (NDD) outcomes and associated risk factors, a Random Forest classification model was employed. The Random Forest algorithm was chosen for its robustness in handling datasets with mixed variable types (categorical and continuous), its ability to handle missing data, and its capacity to rank the importance of predictors, which aligns well with the goals of this analysis.

The response variables comprised 13 NDD outcomes, including ADHD, autism, developmental delay, and speech disorders. The independent variables included 20 key risk factors, such as age (SC_AGE_YEARS), physical activity (PHYSACTIV), screen time (SCREENTIME), housing conditions (housing_22), and other individual, clinical, and family environment variables.

To build and validate the model, the dataset was split into training (70%) and testing (30%) subsets. The model's hyperparameters were optimized to ensure balanced performance across classes, addressing potential imbalances in NDD outcomes. The model was evaluated using confusion matrices and performance metrics like accuracy and kappa.

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) metrics were calculated for each NDD outcome to assess the model's predictive performance further. The ROC-AUC analysis provided an intuitive understanding of the model's sensitivity and specificity for distinguishing between NDD classes, highlighting its effectiveness in predicting specific outcomes. This comprehensive approach ensured that both prediction and variable importance were adequately addressed

Results:

The statistical analysis aimed to identify significant associations between neurodevelopmental disorder (NDD) outcomes and various risk factors. A Chi-Square test was conducted to evaluate the relationships, and a correlation matrix was generated to assess the strength of associations between selected variables.

The Chi-Square analysis revealed significant associations across multiple NDD outcomes and risk factors, with p-values less than 0.05. For example:

- Cerebral palsy (palsy_22) was significantly associated with SC_AGE_YEARS (p = 0.0009), BMI3_22 (p = 0.0035), PHYSACTIV (p = 0.0005), and SCREENTIME (p = 0.0009).
- Seizure disorder (seizure_22) showed strong associations with SC_RACE_R (p = 0.0439), BMI3_22 (p = 0.0005), and PHYSACTIV (p = 0.0005).
- Tourette Syndrome (tourette_22) demonstrated significant relationships with several risk factors, including SC_AGE_YEARS, SC_SEX, and BMI3_22 (all p < 0.001).

The correlation matrix also showed high interdependencies among NDD variables. Strong correlations (r > 0.95) were observed between behavior_22 and depress_22 (r = 0.97), speech_22 and learning_22 (r = 0.98), indicating overlapping behavioral and developmental characteristics among these outcomes.

Model Used

A Random Forest classification model was applied to investigate the relationships between neurodevelopmental disorder (NDD) outcomes and selected risk factors. Random Forest was chosen for its robustness in handling mixed data types, ability to model complex interactions, and feature importance ranking. Thirteen NDD outcomes, such as ADHD, autism, developmental delay, and speech or language disorders, were treated as response variables. The dataset included 20 independent variables (risk factors) such as SC_AGE_YEARS (age), SCREENTIME, BMI3_22, PHYSACTIV, HOURSLEEP, and housing conditions.

Model Evaluation Metrics

The Random Forest model was evaluated using Accuracy, Kappa statistic, and Area Under the Curve (AUC) metrics.

- Overall Accuracy: The model achieved an impressive accuracy of 95.07%, highlighting its predictive capability.

- Kappa Statistic: The Kappa value 0.86 reflects a strong agreement beyond chance, indicating the model's reliability.
- ROC Curve and AUC: The ROC curve was generated for binary classification of DevDelay_22 (Class 1 vs All) to evaluate the model's sensitivity and specificity. The AUC was calculated as 0.976, demonstrating excellent performance distinguishing between participants with developmental delays and those without.

The confusion matrix further supported the model's performance:

- For Class 1 (Developmental Delay), the Sensitivity was 99.02% and the Specificity was 82.49%.
- Other classes, such as Class 3 and Class 99, had lower sensitivities, reflecting imbalances in the dataset but overall strong predictive performance for the primary outcome classes.

Major Findings

The Random Forest model identified significant predictors for NDD outcomes based on Variable Importance. The top-ranked predictors, based on Mean Decrease Accuracy and Mean Decrease Gini scores, included:

1. SC_AGE_YEARS (age): A critical predictor across all outcomes, emphasizing the role of age in NDD risk.
2. speech_22 and learning_22: These variables demonstrated the highest importance, highlighting their association with other neurodevelopmental delays.
3. SCREENTIME: Excessive screen time emerged as a significant environmental risk factor for NDDs.
4. Behavioral problems and genetic conditions were identified as strong predictors for developmental outcomes.
5. HOURSLEEP and bullied_22: Sleep patterns and experiences of bullying also played essential roles, reinforcing the influence of environmental and social stressors on NDDs.

Tables and Figures

Visualizations support key findings:

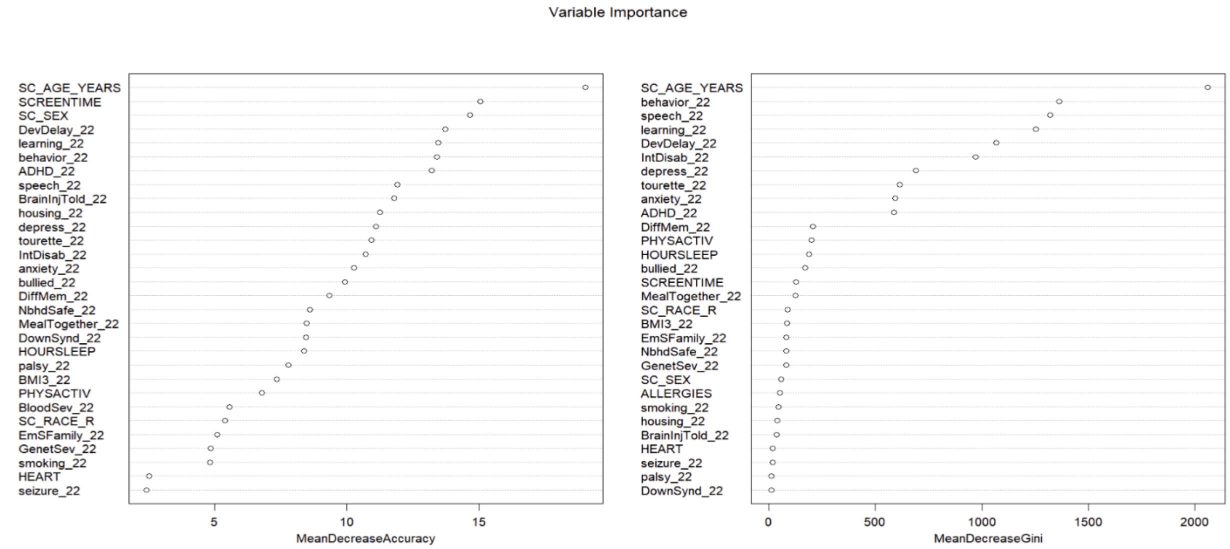

Figure 1 illustrates Variable Importance Plot (Mean Decrease Accuracy and Mean Decrease Gini): The plots demonstrate the ranking of risk factors, with SC_AGE_YEARS, speech_22, and learning_22 emerging as the most influential variables.

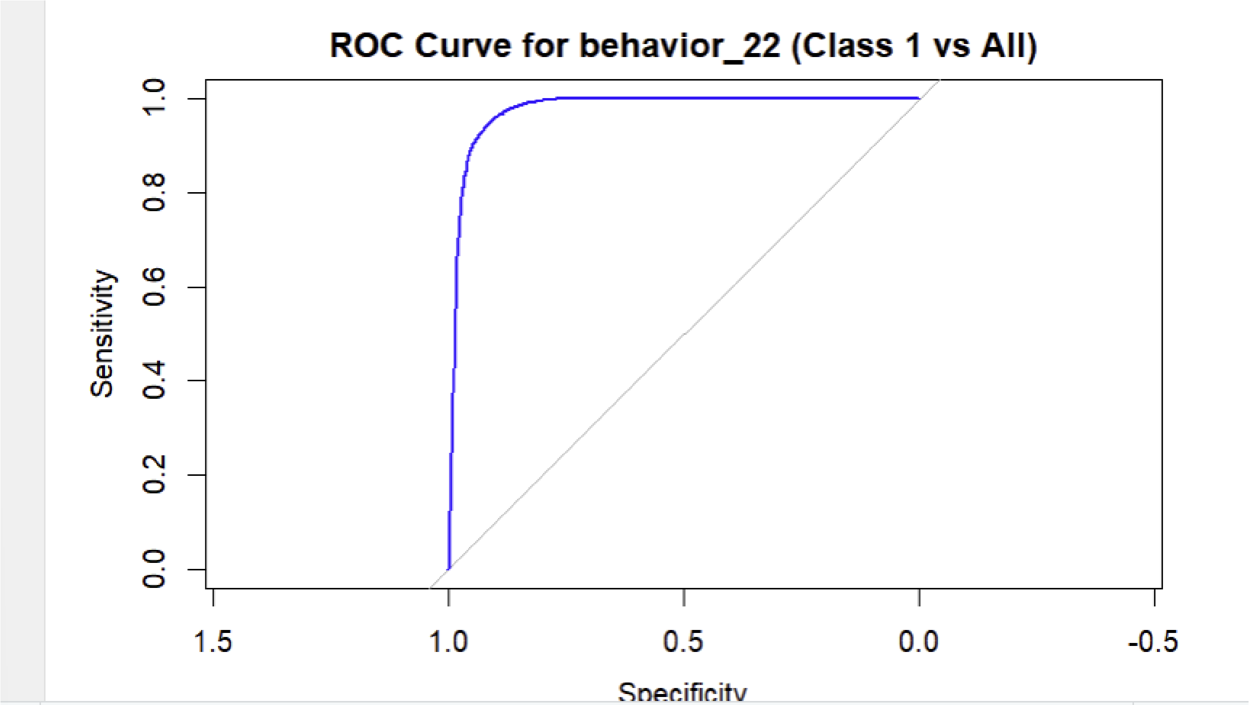

Figure 2 illustrates ROC Curve for DevDelay_22 (Class 1 vs All): The curve illustrates the model's high sensitivity and specificity, with an AUC of 0.976 validating the model's predictive accuracy.

1. Confusion Matrix:
    - The confusion matrix highlights the model's performance, particularly its ability to predict developmental delays (Class 1) with exceptional accuracy and low error rates.

Metrics	Values
Accuracy	0.9507
95% CI	(0.9473, 0.954)
No Information Rate	0.773
P-Value [Acc > NIR]	< 2.2e-16
Kappa	0.8598
Mcnemar's Test P-Value	NA

Table 1 contains information about Confusion Matrix Overall Statistics

Class	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Class : 1	0.9902	0.8249	0.9506	0.9611	0.773	0.7654	0.8052	0.9076
Class : 2	0.04	0.999812	0.785714	0.98372	0.016944	0.000678	0.000863	0.519906
Class : 3	0.56675	0.99013	0.75444	0.97713	0.05077	0.02877	0.03814	0.77844
Class : 95	1	1	1	1	0.1553	0.1553	0.1553	1
Class : 99	0.138462	1	1	0.996548	0.004005	0.000555	0.000555	0.569231

Table 2 contains information about Confusion matrix by Class.

The results indicate that a combination of demographic (age, gender), environmental (screen time, sleep), and clinical variables (speech, behavior, and genetic conditions) significantly influence NDD outcomes. The Random Forest model effectively identified key predictors and demonstrated high accuracy, making it a valuable tool for assessing risk factors and predicting NDD outcomes. These findings underscore the importance of considering multiple factors in the early identification and management of neurodevelopmental disorders.

Discussion

Summary of Methods and Results

This study employed a combination of traditional statistical analyses and advanced machine learning techniques to investigate the associations between **neurodevelopmental disorder (NDD) outcomes** and a set of **20 risk factors**. **Chi-square tests** identified significant relationships between NDD outcomes (e.g., cerebral palsy, autism, developmental delay) and predictors such as **screen time, physical activity, and age**. These findings provided foundational insights into the associations between risk factors and NDDs.

A Random Forest classification model was implemented to further explore these relationships and assess their predictive capabilities. The model demonstrated **high predictive accuracy** (95.07%) and a firm agreement beyond chance, as indicated by a **Kappa value of 0.86**. The **ROC curve analysis** for developmental delay (DevDelay_22) yielded an impressive **AUC of 0.976**, highlighting the model's ability to discriminate between children with and without developmental delays. Variable importance measures revealed **age (SC_AGE_YEARS), screen time (SCREENTIME), speech delay (speech_22), and behavioral problems (behavior_22)** as the top predictors of NDD outcomes.

Strengths

1. **Large, Nationally Representative Dataset**: Utilizing the NSCH 2022 dataset enabled a robust analysis of over 54,000 participants, ensuring the findings' generalizability.
2. **Integration of Machine Learning**: The use of Random Forest allowed for the identification of non-linear relationships and feature importance ranking, providing insights beyond traditional statistical methods.
3. **Identification of Modifiable Risk Factors**: Significant predictors such as **screen time, physical activity, and sleep patterns** underscore actionable areas for intervention.

Limitations

1. **Self-Reported Data**: The reliance on survey responses introduces **recall bias** and potential inaccuracies in reporting clinical conditions and risk factors.
2. **Class Imbalance**: Rare NDD outcomes (e.g., Tourette Syndrome, severe genetic conditions) were underrepresented, which may have affected the model's predictive sensitivity for these classes.
3. **Cross-Sectional Data**: The data represents a single time point, limiting the ability to infer causal relationships between risk factors and NDD outcomes. Longitudinal studies are required to assess causality.

Conclusions

Using advanced machine learning techniques, this study successfully identified key risk factors influencing neurodevelopmental disorders (NDDs) in children aged 3–17. The findings emphasize the importance of **modifiable environmental and lifestyle factors**, such as **screen time, physical activity, and sleep duration**, alongside critical demographic predictors like age. The **Random Forest model** provided high predictive accuracy and robust variable importance rankings, showcasing its potential as a powerful tool for NDD risk assessment.

The results highlight actionable targets for interventions, particularly reducing screen time, promoting physical activity, and encouraging sufficient sleep to mitigate NDD risks. Additionally, age and behavioral delays emerged as primary predictors, suggesting the need for early screening and intervention programs during critical developmental windows.

Future research should focus on:

1. **Longitudinal Data**: To establish causal relationships between risk factors and NDD outcomes.
2. **Intervention Studies**: To test the effectiveness of targeted strategies to reduce modifiable risks.
3. **Balanced Datasets**: Addressing class imbalances for rare NDDs to improve model sensitivity and generalizability.

By leveraging machine learning for NDD prediction, this study offers valuable insights for clinicians, policymakers, and researchers to support early identification and improve outcomes for children at risk of neurodevelopmental disorders.

References:

1. Neurodevelopmental delay: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data. PMCID: PMC6899448.
2. From Neurons to Neighborhoods: The Science of Early Childhood Development. PMID: 25077268
3. Attainment of sitting and walking predicts development of productive vocabulary between ages 16 and 28 months. PMID: 22982273
4. Machine Learning for Predicting Neurodevelopmental Disorders in Children. https://doi.org/10.3390/app14020837