132 (2026) 202865

Contents lists available at ScienceDirect

Research in Autism

journal homepage: www.elsevier.com/locate/rasd

Symptom-based classification of 16p11.2 copy number variations underlying the multidimensional autism spectrum disorder phenotype using machine learning methods

Hilmi Bolat [a,b,*], Gül Ünsel-bolat [c], Edanur Bulut [a], Semiha Özgül [d], Duygu Selin Turan [e], Samet Çelik [f], Özgür Ozan Koyuncu [g], Özgür Kirbiyik [h], Özge Özer Kaya [h], Yaşar Bekir Kutbay [h], Merve Saka Güvenç [h], Kadri Murat Erdoğan [h], S¸ener Arikan [h], Tuba Sözen Türk [h], Altuğ Koç [h], Taha Reşid Özdemir [i], Berk Özyilmaz [i], Gonca Özyurt [j], Burak Ordin [e], Buket Kosova [b,k]

[a] Department of Medical Genetics, Balikesir University Faculty of Medicine, Balikesir, Turkey
[b] Ege University Institute of Health Sciences, Department of Health Bioinformatics, Izmir, Turkey
[c] Department of Child and Adolescent Psychiatry, Balikesir University Faculty of Medicine, Balikesir, Turkey
[d] Department of Biostatistics and Medical Informatics, Ege University Faculty of Medicine, Izmir, Turkey
[e] Faculty of Science, Department of Mathematics, Ege University, Izmir, Turkey
[f] Department of Psychology, Bartin University, Bartin, Turkey
[g] Department of Neuroscience, Institute of Health Sciences, Ege University, Izmir, Turkey
[h] Department of Medical Genetics, Izmir City Hospital, Izmir, Turkey
[i] Department of Medical Genetics, University of Health Sciences, Izmir Faculty of Medicine, Izmir, Turkey
[j] Medical School, Child and Adolescent Psychiatry Department, İzmir Katip Çelebi University, Izmir, Turkey
[k] Department of Medical Biology, Ege University Faculty of Medicine, Izmir, Turkey

ARTICLE INFO

Keywords:
16p11.2
CNVs
Machine Learning
Neurodevelopmental Disorders

ABSTRACT

Purpose: Copy number variations (CNVs) in the 16p11.2 region are well-established contributors to neurodevelopmental disorders, yet phenotype variability across this locus remains insufficiently characterized. This study investigates clinical features and ASD-related symptoms among carriers of rare pathogenic and common CNVs, and evaluates symptom-level discriminability using machine learning (ML) methods.

Methods: Genetic data from 7568 individuals were retrospectively screened, identifying 147 carriers of 16p11.2 CNVs. Detailed clinical assessments were completed for 50 participants. ASD-related symptoms were evaluated using a structured 25-item instrument. Group comparisons applied nonparametric statistics with effect sizes, confidence intervals, and FDR correction. ML analyses used PCA and k-means for feature selection, oversampling methods (SMOTE, Borderline-SMOTE, ADASYN), and five classifiers, evaluated through cross-validation.

Results: Across pathogenic and common CNV groups, no significant differences were observed in social communication, restricted/repetitive behaviors, sensory symptoms, regression, or total autism scores (FDR-adjusted $p > 0.05$). Aggression was more frequently endorsed in pathogenic

* Correspondence to: BOLAT. Balikesir University, Faculty of Medicine, Department of Medical Genetics, Cagıs Campus, Balikesir-Altieylül, Turkey.
E-mail address: hilmi_bolat@hotmail.com (H. Bolat).

https://doi.org/10.1016/j.reia.2026.202865
Received 26 August 2025; Received in revised form 3 January 2026; Accepted 12 February 2026
Available online 19 February 2026
3050-6565/© 2026 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

CNV carriers (raw p = 0.030; FDR p = 0.098). BMI was higher in pathogenic CNVs, though nonsignificant after correction (raw p = 0.027; FDR p = 0.152). ML analyses identified three recurrent discriminative symptoms across multiple datasets: delayed response to name, unusual object play, and aggression. Dataset 3 (16 symptoms) provided the most balanced classification performance but, given the very small pathogenic CNV sample, results remain exploratory.

Conclusion: Findings suggest that, while most autism-related symptoms do not differ between groups, aggression and increased BMI may represent preliminary phenotypic signals associated with pathogenic CNVs. Integrating clinical data from 147 CNV carriers further supports a potential widespread effect across the broader 16p11.2 locus rather than a single breakpoint-specific mechanism. However, all results should be interpreted cautiously due to limited sample size, and larger, systematically phenotyped cohorts are required to establish robust genotype–phenotype relationships.

1. Introduction

Genetic factors are emphasized as important contributors in the etiology of intellectual disability (ID), attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD) (Sicherman et al., 2021). By the improvement of genetic methods, copy number variations (CNVs) and related genes associated with neurodevelopmental disorders are being elucidated using techniques such as microarray analysis or whole exome/genome sequencing. CNVs are defined as structural changes that involve both duplications and deletions of DNA sequences. CNVs vary in terms of size, gene content and prevalence (Bijlsma et al., 2009). Each CNV provides information about the different clinical manifestations of many disorders, such as psychiatric disorders, neurological disorders or cancer.

In recent years, studies investigating the genetic-first approach have been implicated for the diagnosis and clinical follow-up of neurodevelopmental disorders. It has been shown that a genetic-first approach may provide more power to detect biomarkers in these disorders (Fetit et al., 2020). Thus, it is aimed to increase the specificity of heterogeneous neurodevelopmental disorders. One of the leading research areas in this approach has been CNV research.

So far, 16p11.2 region has been emphasized in CNV studies of neurodevelopmental disorders. However, the clinical relevance of CNVs in this region has not been clearly demonstrated. While 16p11.2 microdeletion is approximately 1/100 in cases diagnosed with ASD, this rate was found in approximately 3 out of 10,000 people in the general population (Medland et al., 2022, Bassuk et al., 2013).

CNVs may cause variable clinical features due to different characteristics such as gene content and size. Therefore, it is difficult to establish genotype-phenotype correlation due to the different contents of these CNVs located in the 16p11.2 region. In 16p11.2 CNV carriers, differences in clinical presentations make accurate clinical interpretation difficult and often lead to delayed diagnosis of neuropsychiatric symptoms. To better understand this relationship, further research is needed to investigate the neurodevelopmental and neuropsychiatric outcomes of 16p11.2 carriers to provide evidence-based clinical care. It is important to address and treat symptoms associated with CNVs in the 16p11.2 region early to achieve functional and physical health outcomes (Mitchell, 2011).

In recent studies, it has been emphasized that a systematic and standardized evaluation and subgrouping of symptoms, not only in terms of diagnosis, should be performed to ensure genetic homogeneity of CNVs that are different from each other in terms of content in the same region (Bassuk et al., 2013). Another issue emphasized as a limitation in the studies is that the studies are limited to a single

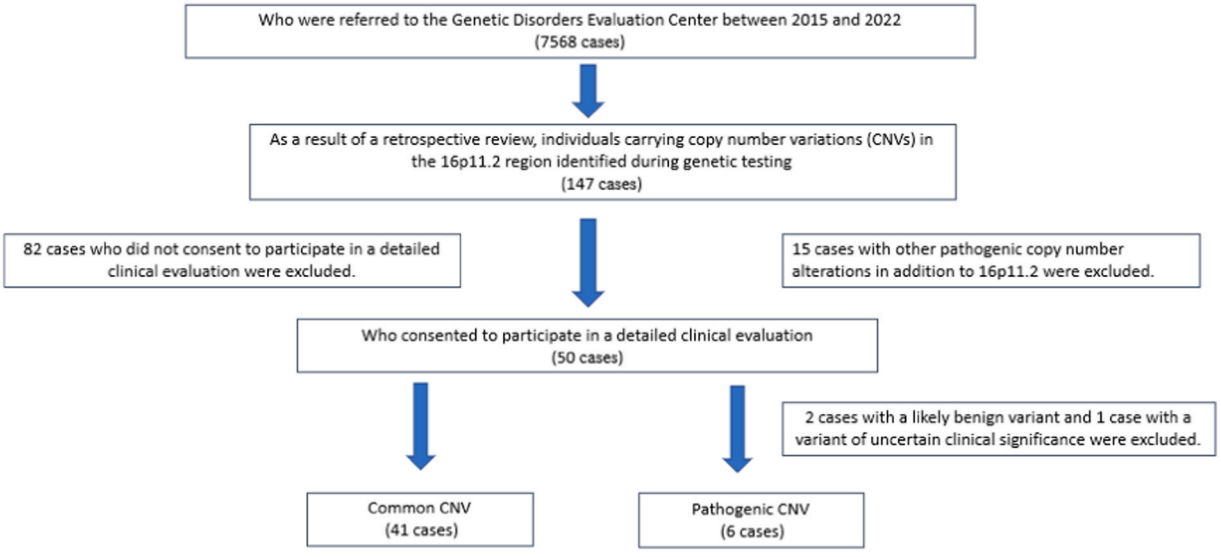

Fig. 1. Flowchart of patient selection and inclusion in the study.

diagnostic group such as ASD. By examining all individuals with this CNV, with or without ASD diagnosis, and including all cases for the general population, the risks related to CNV regions can be revealed more clearly and accurately (Niarchou et al., 2019).

In our study, clinical data of individuals with rare pathogenic and common CNVs belonging to the 16p11.2 region were obtained and detailed psychiatric evaluations were made. We evaluated the association between common and rare variations of the 16p11.2 region CNVs and neurodevelopmental disorders through a detailed clinical interview.

2. Materials and methods

2.1. Sample

In the study, the genetic results of 7568 patients who applied to the genetic diseases evaluation centre between 2015 and 2022 were retrospectively screened. Among these patients, patients with copy number alterations in the 16p11.2 region were selected and invited to participate in the study. As a result of the retrospective file review, 147 cases (147/7568, 1.94 %) were found to have 16p11.2 CNVs. The diagnoses and indications for testing in 147 cases are presented in Supplementary Table 1. All sampling procedures were presented in Fig. 1.

In the design of the study, clinical evaluation of cases with CNV in the 16p11.2 region as a result of genetic analyses was performed in terms of neurodevelopmental disorders. Ethics committee approval (Ege University Medical Research Ethics Committee - approval decision 21–9 T/10) was obtained for the study.

2.2. CNV classification

CNVs were classified as pathogenic or common according to established clinical interpretation frameworks. Pathogenic CNVs were defined as rare copy number changes overlapping known disease-associated regions within 16p11.2, characterized by low population frequency, substantial gene content, and prior associations with neurodevelopmental or neuropsychiatric phenotypes in public databases such as DECIPHER. In contrast, common CNVs were defined as recurrent variants within the 16p11.2 locus that are more frequently observed in the general population, typically lack consistent disease association, and are often reported as benign or low-penetrance variants in clinical databases. CNVs were classified as pathogenic/likely pathogenic or common/likely benign according to ACMG/ClinGen criteria, integrating size, gene content, dosage sensitivity, population frequency, and known disease associations.

The list of genes encompassed by pathogenic and common CNVs, together with the start and end coordinates of the copy number changes, is provided in Supplementary Table 2.

Table 1
Autism spectrum and associated symptoms.

Social and Communication
A1. Delayed speech
A2. Delayed response to name
A3. Poor eye contact
A4. Lack of gestures (e.g., pointing, nodding or shaking head)
A5. Difficulty understanding gestures
A6. Preference to play alone or play with objects rather than with others
A7. More focus on objects than people
A8. Failure to initiate or respond to social interactions
A9. Difficulties in initiating and/or maintaining relationships and friendships
Restricted and Repetitive behaviors
A10. Played with toys or objects in an unusual way (e.g. repetitive play, lining up toys)
A11. Need for sameness (e.g. difficulties with changes in routine)
A12. Unusual motor mannerisms (e.g. hand flapping, spinning
A13. Unusual interest in specific objects or toys (e.g. high in intensity or focus)
Sensory Symptoms
A14. Unusual interest in sensory aspects of the environment (e.g. excessive smelling of objects or people, fascination with lights or movement)
A15. Sensory hyperreactivity (e.g. excessive or adverse response to specific sounds, lights, touch, smell or tastes)
A16. Sensory hyporeactivity (e.g. insensitivity or indifference to sensory pain or temperature, slow response to sensory stimuli in the environment)
Aggression
A17.Temper tantrums
A18.Aggression toward self
A19.Aggression toward others
Regression
A20.Loss of skills
A21.Loss of language (words only)
A22.Loss of language (phrases)
A23.Less social engagement
A24.Loss of motor skills
A25.Loss of daily living skills

Adapted by Sicherman et al., (2021).

2.3. Statistical analyses

Descriptive statistics included frequencies and percentages for categorical variables, and mean, standard deviation (SD), median, and range (minimum-maximum) for numerical variables. Comparisons of numerical demographic data between groups were performed using independent sample T-test or Mann-Whitney *U* test (Wilcoxon rank sum) for non-normally distributed data. Categorical data were compared using Pearson Chi-square test or Fisher's exact test when expected frequencies were low. Statistical significance was set at $p < 0.05$. Multiple comparisons were corrected using the Benjamini-Hochberg (FDR) method. Effect sizes were reported as Cohen's d or Mann-Whitney r for numerical variables, and Cramer's V for categorical variables. All statistical analyses were conducted using R (R Foundation for Statistical Computing, Vienna, Austria; version 4.0.5, arsenal package) and Python (version 3.10 +) programming languages.

2.4. Machine learning

In the machine learning phase of the study, after data preprocessing, PCA (Jolliffe, 2011) and the k-means clustering (MacQueen, 1967) algorithms were used for feature selection. Subsequently, the dataset was balanced using SMOTE (Chawla et al., 2002), Borderline-SMOTE (Han et al., 2005), and ADASYN (He et al., 2008) for oversampling. During the classification phase, C4.5 decision tree classifier (Quinlan, 2014), Random Tree classifier (Breiman et al., 1984), Random Forest classifier (Breiman, 2001), Logistic Regression (Hosmer, 2013), and Support Vector Machines (Cortes & Vapnik, 1995) algorithms were employed. Finally, the obtained results were evaluated using accuracy, precision, recall, and F-measure metrics. Detailed symptom screening and scales were used in our study for the diagnosis of ASD which is closely associated with this CNV. Social communication domain (9 items), restricted and repetitive behaviors (4 items), sensory symptoms (3 items), aggression (3 items), and regression (6 items) were evaluated.

Symptom Assessment

The symptoms listed in Table 1 were adapted from the study conducted by Sicherman et al. (2021), which investigated early behavioural signs that are stronger predictors of ASD and related conditions. In their work, parents were provided with a structured list of 25 clinical signs, derived primarily from DSM-based diagnostic criteria for ASD, with additional items incorporated following a pilot study in which parents identified commonly observed behaviours that prompted diagnostic referral. Each symptom was reported using a binary format, a scoring structure that we retained in the present study. This instrument has demonstrated ecological validity and clinical relevance for characterising early ASD-related behaviours, and its established reliability supports its use in exploratory analyses examining behavioural variation across 16p11.2 CNV subgroups. While not originally developed for CNV-specific phenotyping, the dataset provides a standardised framework for assessing symptom patterns that may differentiate clinically significant CNVs from common variants.

Table 2
Sociodemographic Characteristics and Anthropometric Measurements.

	Common CNV	Pathogenic CNV	Total	***raw_p*** value	Effect Size	***FDR_p$_{adj}$***	Confidence Interval (95 % CI)
Age (n)	41	6	47	0.9741[1]	0.0334	1	40.0 – 47.4
Mean (sd)	16.5 (14.3)	11.5 (5.4)	15.9 (13.6)				
Median (min, max)	9.0 (3.0, 49.0)	10.0 (6.0, 18.0)	9.0 (3.0, 49.0)				
Sex (n)	41	6	47	0.2042[2]	0.3462	0.756	
Male	21 (%51.2)	5 (%83.3)	26 (%55.3)				
Female	20 (%48.8)	1 (%16.7)	21 (%44.7)				
Mother Age (n)	39	6	45	0.3321[1]	0.2258	0.5507	40.1 – 47.3
Mean (sd)	44.5 (12.8)	38.8 (7.8)	43.7 (12.4)				
Median (min, max)	40.0 (28.0, 74.0)	37.5 (30.0, 50.0)	39.0 (28.0, 74.0)				
Father Age (n)	35	6	41	0.4941[1]	0.1923	0.5602	41.5 – 47.5
Mean (sd)	44.9 (9.4)	41.8 (8.2)	44.5 (9.2)				
Median (min, max)	43.0 (32.0, 72.0)	41.5 (32.0, 55.0)	43.0 (32.0, 72.0)				
Height (n)	30	3	33	0.3011[1]	-0.6679	0.5507	133.4 – 155.6
Mean (sd)	142.4 (31.9)	165.3 (13.1)	144.5 (31.2)				
Median (min, max)	140.5 (62.0, 195.0)	161.0 (155.0, 180.0)	155.0 (62.0, 195.0)				
Weight (n)	38	6	44	0.1281[1]	-1.2458	0.2261	36.7 – 54.3
Mean (sd)	43.3 (28.6)	59.7 (29.6)	45.5 (29.0)				
Median (min, max)	35.0 (5.0, 96.0)	52.5 (24.0, 110.0)	37.5 (5.0, 110.0)				
BMI (n)	28	3	31	*0.0271*[1]	*-1.6915*	*0.1522*	18.6 – 23.0
Mean (sd)	19.9 (5.5)	28.9 (5.0)	20.8 (6.0)				
Median (min, max)	19.1 (11.9, 31.3)	28.5 (24.1, 34.0)	20.7 (11.9, 34.0)				

1Wilcoxon rank sum test, 2 Fisher's Exact Test for Count Data

3. Results

3.1. Classification of copy number variations

Among the 147 cases included in Supplementary Table 1, 12 (8.16 %) were classified as pathogenic and 6 (4.08 %) as likely pathogenic, yielding a total of 18 CNVs, which is 12.24 % of our cohort. Among the 18 cases classified as pathogenic or likely pathogenic, 12 cases presented with neuropsychiatric symptoms, most commonly intellectual disability, developmental delay, epilepsy, ASD, hypotonia, and dysmorphic features. Across the 18 pathogenic and likely pathogenic CNVs, genomic breakpoints clustered within the 16p11.2 region, spanning approximately from 2833117 to 30365165 (hg19), with variant sizes ranging from about 460 kb to 1.3 Mb and encompassing the OMIM morbid genes, most notably *TBX6, PRRT2, KIF22, ALDOA, MAPK3, MVP, DOC2A, SEZ6L2, TAOK2, HIRIP3, CORO1A*, and *QPRT* genes.

Of these 109 common CNVs, 79 cases (72.5 %) were referred with at least one neuropsychiatric symptom, most frequently intellectual disability, developmental delay, and ASD. The breakpoints of common CNVs were widely distributed across the 16p11.2 locus, extending from 31,935,366–33,863,672 (hg19), and the CNV sizes ranged from roughly 180 kb to 1.9 Mb. These common CNVs encompassed OMIM gene including *TP53TG3*.

3.2. Statistical analysis results

According to the results in Table 2, there were no statistically significant differences between the common CNV and pathogenic CNV groups in terms of age, sex, or parental ages ($raw\_p > 0.05, FDR\_p_{adj} > 0.05$). Height and weight also did not differ significantly between the groups ($raw\_p > 0.05, FDR\_p_{adj} > 0.05$). Although the median Body Mass Index (BMI) was higher in the pathogenic CNV group (28.5) compared to the common CNV group (19.1), the difference was not statistically significant after multiple comparison correction ($raw_p = 0.027, FDR\_p_{adj} = 0.152$). The effect size was large for BMI, while it was small to moderate for the other demographic and anthropometric variables.

Table 3 presents the developmental stages of the patients. Regarding speech, 90.2 % of the common CNV group and 100 % of the pathogenic CNV group had acquired speech, with no significant difference between the groups ($raw\_p = 1, FDR\_p_{adj} = 1$). The median age for first word use was 18 months in the common CNV group and 21 months in the pathogenic CNV group, which was not statistically significant ($raw\_p = 0.455, FDR\_p_{adj} = 0.582$). Similarly, there was no significant difference in independent walking; the median age for walking was 18 months in the common CNV group and 24 months in the pathogenic CNV group ($raw\_p = 0.351$, $FDR\_p_{adj} = 0.560$). Overall, no significant differences were observed between the common and pathogenic CNV groups regarding developmental milestones.

Table 4 presents the past medical history and family history of the patients. No statistically significant differences were observed between the common CNV and pathogenic CNV groups regarding psychiatric medication use, history of epilepsy, parental consanguinity, presence of similar diseases in the family, or family history of neuropsychiatric disorders ($raw\_p > 0.05, FDR\_p_{adj} > 0.05$). These findings indicate that the two groups were generally similar in terms of medical and family history.

Table 5 presents the statistical analysis of autism-related symptoms in patients with common and pathogenic CNVs. No significant differences were observed between the groups in social communication, restricted and repetitive behaviors, or sensory symptoms ($raw\_p > 0.05, FDR\_p_{adj} > 0.05$). Regression scores were also similar between the groups ($p > 0.05$). However, aggression symptoms were significantly higher in the pathogenic CNV group compared to the common CNV group ($raw\_p = 0.030, FDR\_p_{adj} = 0.098$). Overall total autism scores did not differ significantly between the groups ($p = 0.251$).

Table 3
Developmental Stages.

	Common CNVs	Pathogenic CNVs	Total	*raw_p* value	Effect Size	*FDR_p_adj*	Confidence Interval (95 % CI)
Speech (n)	41	6	47	1.000[a]	0	1	
Yes	37 (%90.2)	6 (%100.0)	43 (%91.5)				
No	4 (%9.8)	0 (%0.0)	4 (%8.5)				
First word use (month) (n)	37	6	41	0.455[2]	0.1672	0.5817	18.6 – 27.4
Mean (sd)	22.3 (13.8)	27.5 (18.2)	23.0 (14.3)				
Median (Min, max)	18.0 (8.0, 60.0)	21.0 (9.0, 60.0)	18.0 (8.0, 60.0)				
Independent Walking (n)	41	6	47	0.571[a]	0	1	
Yes	34 (%82.9	6 (%100.0)	40 (%85.1)				
No	7 (%17.1)	0 (%0.0)	7 (%14.9)				
First Walking time (month) (n)	34	6	40	0.351[2]	0.1840	0.5602	17.4 – 25.8
Mean (sd)	21.3 (14.1)	23.0 (9.6)	21.6 (13.4)				
Median (Min, max)	18.0 (10.0, 72.0)	24.0 (12.0, 36.0)	18.0 (10.0, 72.0)				

[a] Wilcoxon rank sum test, [2] Fisher's Exact Test for Count Data

Table 4
Past Medical History and Family History of Cases.

	Common CNVs	Pathogenic CNVs	Total	*raw_p*value	Effect Size	*FDR_p*$_{adj}$	Significant
Psychiatric medication use(n)	41	6	47	1.000[1]	0.4444	0.6725	False
Yes	5 (%12.2)	1 (%16.7)	6 (%12.8)				
No	36 (%87.8)	5 (%83.3)	41(%87.2)				
Epilepsy(n)	39	5	44	0.173[1]	0.2	0.5507	False
Yes	5 (%12.8)	2 (%40.0)	7 (%15.9)				
No	34 (%87.2)	3 (%60.0)	37(%84.1)				
Parental consanguinity (n)	41	5	46	0.594[1]	0.5882	1	False
Yes	10 (%24.4)	2 (%40.0)	12 (%26.1)				
No	31 (%75.6)	3 (%60.0)	34 (%73.9)				
Similar disease in family(n)				0.248[1]	0.3158	0.5817	False
Yes	7 (%17.1)	2 (%40.0)	9 (%19.6)				
No	34 (%82.9)	3 (%60.0)	37 (%80.4)				
Family history of neuropsychiatric disease(n)	41	5		1.000[1]	∞	1	False
Yes	10 (%24.4)	1 (%20.0)	11 (%23.9)				
No	31 (%75.6)	4 (%80.0)	35 (%76.1)				

[1]Fisher's Exact Test for Count Data

3.3. Machine learning analysis results

The machine learning section is organized sequentially as data preprocessing, feature selection, oversampling, classification, and evaluation. After data preprocessing, Two different feature selection techniques from machine learning methods were used to determine which of the 25 different Autism symptoms examined under five main headings are more effective in distinguishing patients with pathogenic CNV and common CNV (Table 6). One of the methods was Principal Component Analysis (PCA), while in the second method, feature selection was carried out using the k-means clustering algorithm. When PCA is applied to the dataset and the classification process is performed, the autism symptoms in the resulting decision tree are 5 symptoms (DS2) (Table 7). When the k-means algorithm was used for feature selection, the 25 autism symptoms were grouped according to their similarities. The grouping process started from two clusters up to 23 clusters. For each clustering result, the symptoms representing the clusters were decided by determining the autism symptom closest to the center of each cluster. Classification operations were performed with the representative autism symptoms, and accuracy rates were obtained. The results obtained are given in Table 6. When the representative symptoms in the results were combined, 16 representative Autism symptoms (DS3) were obtained.

Although the number of symptoms differs across the three datasets, the number of samples in the pathogenic CNV and common CNV classes is 6 and 41, respectively, indicating that the dataset is imbalanced. While learning algorithms can effectively learn the common CNV class, they struggle to learn the pathogenic CNV class. To address this learning difficulty, the dataset was first balanced using three oversampling algorithms: SMOTE (Synthetic Minority Over-sampling Technique), Borderline SMOTE (Borderline Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning). The balanced datasets were then classified using five different classifiers. The classifiers used are: J48 (C4.5 decision tree classifier), RT (Random Tree classifier), RF (Random Forest classifier), Logistic Regression, and SMO (Support Vector Machine). Model performance was evaluated using 5-fold cross-validation. The performance of the classifiers was compared using four different performance metrics. The performance metrics used are: Accuracy, Precision, Recall, and F-measure.

Table 8 presents the classification performance metrics for datasets balanced using SMOTE. Examination of the results indicates that the DS3 dataset is the most suitable for classification, as it consistently achieves the highest and most balanced values across accuracy, recall, and F-measure. While DS1 also demonstrates satisfactory performance, particularly for algorithms other than Logistic Regression, it lacks the overall consistency observed in DS3. Although DS2 attains high precision in certain algorithms, its comparatively lower recall and accuracy reduce its reliability. Therefore, based on the results in Table 8, DS3 can be considered the most balanced and effective dataset.

Table 9 presents the classification performance metrics for datasets balanced using Borderline SMOTE. DS3 achieves the highest metric values for all classifiers except Logistic Regression. Although DS2 attains high precision in some cases, its lower recall and accuracy make it less reliable. Overall, according to Table 9, DS3 can be considered the most balanced and effective dataset for most algorithms.

Table 10 presents the classification performance metrics for datasets balanced using ADASYN. Examination of the results reveals similar outcomes to those in Table 9. DS3 achieves the highest metric values for all classifiers except Logistic Regression.

Across all three oversampling methods—SMOTE, Borderline SMOTE, and ADASYN—DS3 generally demonstrates the highest and most balanced classification performance for most algorithms. While DS1 and DS2 show strengths in specific metrics or classifiers, DS3 emerges as the most reliable and effective dataset overall; therefore, it can be inferred that the existing symptoms in DS3 are more discriminative in distinguishing pathogenic and common CNVs.

Table 5
Statistical analysis of Autism symptoms.

	Common CNVs	Pathogenic CNVs	Total	*raw_p*value	Effect Size	FDR_p_{adj}	Confidence Interval (95 % CI)
A_1 (n)	41	6	47	1.000[1]	6.8	0.4158	
0	23 (%56.1)	4 (%66.7)	27 (%57.4)				
1	18 (%43.9)	2 (%33.3)	20 (%42.6)				
A_2 (n)	41	6	47	1.000[1]	0	1	
0	35 (%85.4)	6 (%100.0)	41 (%87.2)				
1	6 (%14.6)	0 (%0.0)	6 (%12.8)				
A_3 (n)	41	6	47	0.267[1]	10.5	0.5111	
0	35 (%85.4)	4 (%66.7)	39 (%83.0)				
1	6 (%14.6)	2 (%33.3)	8 (%17.0)				
A_4 (n)	41	6	47	1.000[1]	9	0.3536	
0	28 (%68.3)	4 (%66.7)	32 (%68.1)				
1	13 (%31.7)	2 (%33.3)	15 (%31.9)				
A_5	41	6	47	0.637[1]	20	0.2261	
0	31 (%75.6)	4 (%66.7)	35 (%74.5)				
1	10 (%24.4)	2 (%33.3)	12 (%25.5)				
A_6 (n)	41	6	47	0.662[1]	∞	0.1522	
0	26 (%63.4)	3 (%50.0)	29 (%61.7)				
1	15 (%36.6)	3 (%50.0)	18 (%38.3)				
A_7 (n)	41	41		0.326[1]	12.6667	0.2809	
0	31 (%75.6)	3 (%50.0)	34 (%72.3)				
1	10 (%24.4)	3 (%50.0)	13 (%27.7)				
A_8 (n)				0.379[1]	∞	0.1522	
0	25 (%61.0)	2 (%33.3)	27 (%57.4)				
1	16 (%39.0)	4 (%66.7)	20 (%42.6)				
A_9 (n)				0.379[1]	∞	0.1522	
0	25 (%61.0)	2 (%33.3)	27 (%57.4)				
1	16 (%39.0)	4 (%66.7)	20 (%42.6)				
Social communication total score				0.527[2]	0.7278	0.0978	
Mean (sd)	2.7 (3.3)	3.7 (2.9)	2.8 (3.2)				
Median (Min, Max)	1.0 (0.0, 9.0)	5.0 (0.0, 6.0)	1.0 (0.0, 9.0)				
A_10(n)	41	6		0.164[1]	42	0.1311	
0	32(%78.0)	3(%50.0)	35 (%74.5)				
1	9(%22.0)	3(%50.0)	12(%25.5)				
A_11(n)	41	6		0.594[1]	42	0.1311	
0	33(%80.5)	4 (%66.7)	37(%78.7)				
1	8(%19.5)	2 (%33.3)	10(%21.3)				
A_12(n)	41	6		0.326[1]	12.6667	0.2417	
0	31 (%75.6)	3(%50.0)	34 (%72.3)				
1	10 (%24.4)	3 (%50.0)	13 (%27.7)				
A_13(n)	41	6		0.594[1]	10.5	0.45	
0	33(%80.5)	4 (%66.7)	37(%78.7)				
1	8(%19.5)	2 (%33.3)	10(%21.3)				
Restricted and repetitive behaviors (n)	41	6		0.349[2]	0.5839	0.0978	0.6 – 1.4
Mean (sd)	0.9(1.3)	1.7(1.9)	1.0(1.4)				
Median (Min, Max)	0.0(0.0, 4.0)	1.5(0.0, 4.0)	0.0(0.0, 4.0)				
A_14(n)				0.214[1]	20	0.1818	
0	36(%87.8)	4 (%66.7)	40(%85.1)				
1	5(%12.2)	2 (%33.3)	7(%14.9)				
A_15(n)				1.000[1]	3.5	0.6793	
0	23 (%56.1)	4 (%66.7)	27 (%57.4)				
1	18 (%43.9)	2 (%33.3)	20 (%42.6)				
A_16(n)				1.000[1]	10.5	0.45	
0	36(%87.8)	5(%83.3)	41 (%87.2)				
1	5(%12.2)	1(%16.7)	6 (%12.8)				
Sensory symptoms(n)				0.671[2]	0.3909	0.1818	0.4 – 1.0
Mean (sd)	0.7(0.9)	0.8(1.0)	0.7(0.9)				
Median (Min, Max)	0.0(0.0, 3.0)	0.5(0.0, 2.0)	0.0(0.0, 3.0)				
A_17(n)	41	6	47	0.004[1]	42	0.1311	
0	33(%80.5)	1(%16.7)	34 (%72.3)				
1	8(%19.5)	5(%83.3)	13 (%27.7)				
A_18(n)	41	6	47	1.000[1]	5	0.5127	

(*continued on next page*)

Table 5 (*continued*)

	Common CNVs	Pathogenic CNVs	Total	*raw_p*value	Effect Size	*FDR_p*$_{adj}$	Confidence Interval (95 % CI)
0	34(%82.9)	5(%83.3)	39 (%83.0)				
1	7(%17.1)	1(%16.7)	8 (%17.0)				
A_19(n)	41	6	47	1.000[1]	0	1	
0	35 (%85.4)	6 (%100.0)	41 (%87.2)				
1	6 (%14.6)	0 (%0.0)	6 (%12.8)				
Aggression	41	6	47	0.030[2]	0.5156	0.0978	0.3 – 0.9
Mean (sd)	0.5(1.0)	1.0(0.6)	0.6(0.9)				
Median (Min, Max)	0.0(0.0, 3.0)	1.0(0.0, 2.0)	0.0(0.0, 3.0)				
A_20 (n)	40	6	46	1.000[1]	5	0.5127	
0	31 (%77.5)	5 (%83.3)	36 (%78.3)				
1	9 (%22.5)	1 (%16.7)	10 (%21.7)				
A_21 (n)	47	6	47	1.000[1]	3.1667	0.5582	
0	31 (%75.6)	5 (%83.3)	36 (%76.6)				
1	10 (%24.4)	1 (%16.7)	11 (%23.4)				
A_22 (n)	47	6	47	0.579[1]	0	1	
0	32 (%78.0)	6 (%100.0)	38 (%80.9)				
1	9 (%22.0)	0 (%0.0)	9 (%19.1)				
A_23 (n)	47	6	47	0.395[1]	9	0.3141	
0	28 (%68.3)	3 (%50.0)	31 (%66.0)				
1	13 (%31.7)	3 (%50.0)	16 (%34.0)				
A_24 (n)	47	6	47	0.344[1]	12.6667	0.2417	
0	30 (%73.2)	3 (%50.0)	33 (%70.2)				
1	11 (%26.8)	3 (%50.0)	14 (%29.8)				
A_25 (n)	47	6	47	1.000[1]	5	0.5127	
0	31 (%75.6)	5 (%83.3)	36 (%76.6)				
1	10 (%24.4)	1 (%16.7)	11 (%23.4)				
Regression (n)	47	6	47	0.729[2]			0.8 – 2.2
Mean (sd)	1.5 (2.4)	1.5 (2.0)	1.5 (2.3)				
Median (Min, Max)	0.0 (0.0, 6.0)	1.0 (0.0, 5.0)	0.0 (0.0, 6.0)				
	Common CNVs (N = 41)	Pathogenic CNVs (N = 6)	Total (N = 47)	p value			
Total				0.251[1]			4.3 – 8.9
Mean (Sd)	6.2 (8.0)	8.7 (7.2)	6.6 (7.9)				
Median (Min, Max)	2.0 (0.0, 25.0)	8.5 (1.0, 19.0)	2.0 (0.0, 25.0)				

1Wilcoxon rank sum test

4. Discussion

In this study, clinical data from individuals with rare pathogenic and common CNVs in the 16p11.2 region were evaluated, and ASD-related symptoms were examined. These assessments were complemented by biostatistical and bioinformatic analyses. Although detailed clinical analyses were performed on 50 individuals, the study also incorporated a broader dataset of 147 CNV carriers identified during genetic screening, allowing the clinical findings to be contextualized within a larger genomic framework.

In addition to the clinical comparisons, we incorporated genomic data from the full cohort of 147 individuals carrying CNVs in the 16p11.2 region. Among these, 18 CNVs (12.24 %) were classified as pathogenic or likely pathogenic; most carriers exhibited neuropsychiatric features such as intellectual disability, developmental delay, epilepsy, ASD, hypotonia, and dysmorphic findings. And, 109 CNVs were categorized as common variants; however, 72.5 % of these cases also presented with neurodevelopmental symptoms, most frequently intellectual disability, developmental delay, and ASD. Collectively, these genomic findings support the interpretation that phenotypic variability may arise from broader structural variation across the 16p11.2 region rather than being restricted to the classical BP4–BP5 interval.

Biostatistical comparisons showed that only aggression scores were nominally higher in the pathogenic CNV group. Given the imbalance between groups, these comparisons should be interpreted with caution, as estimates, particularly in the pathogenic CNV group, may be unstable and influenced by single cases. Although aggression appeared more frequent among pathogenic CNV carriers, these observations should be considered suggestive and require confirmation in larger, independent cohorts.

Our results are consistent with the study of Weiner et al., in which the authors evaluated the effect of both rare pathogenic and common polygenic variations of the 16p11.2 region on the diagnosis of ASD (Weiner et al., 2022). They showed that both multiple genes and other non-gene included regions associated with 16p11.2 CNVs are effective jointly. They stated that the genetic influence in the diagnosis of ASD is due to both the well-defined 16p11.2 syndrome and the widespread polygenic variations of 16p11.2 region. Both of these effects have been associated with decreased function of cortically expressed genes at 16p. They proposed that a large number of genes in 16p that are specifically expressed in the cortex increase the diagnosis of ASD through these transcriptional changes. As a conclusion, we suggest there is a widespread effect along the 16p11.2 locus for ASD diagnosis that is not limited to a

Table 6
The Clustering Analysis Results.

k (Number of clusters)	K representative symptom	Accuracy obtained from k symptom and j48 classifier
2	A_8, A_25	87.23
3	A_8, A_10, A_25	87.23
4	A_2, A_5, A_10, A_15	87.23
5	A_5, A_22, A_2, A_3, A_10	87.23
6	A_1, A_25, A_4, A_8, A_15, A_7	87.23
7	A_15, A_5, A_16, A_19, A_7, A_8, A_4	87.23
8	A_8, A_10, A_2, A_15, A_16, A_18, A_5, A_7	87.23
9	A_8, A_15, A_16, A_4, A_2, A_5, A_6, A_23, A_18	87.23
10	A_12, A_20, A_3, A_14, A_10, A_8, A_4, A_22, A_1, A_6	80.85
11	A_14, A_21, A_20, A_8, A_6, A_2, A_3, A_7, A_15, A_4, A_12	80.85
12	A_4, A_12, A_5, A_6, A_18, A_13, A_15, A_23, A_8, A_2, A_11, A_1	80.85
13	A_10, A_12, A_14, A_3, A_17, A_18, A_25, A_16, A_13, A_8, A_11, A_15, A_4	74.47
14	A_11, A_2, A_8, A_3, A_1, A_19, A_18, A_5, A_13, A_12, A_14, A_6, A_15, A_21	80.85
15	A_5, A_15, A_22, A_12, A_20, A_2, A_17, A_1, A_21, A_13, A_18, A_6, A_7, A_8, A_3	85.11
16	A_19, A_14, A_23, A_7, A_20, A_8, A_10, A_3, A_21, A_12, A_15, A_17, A_4, A_13, A_1, A_5	78.72
17	A_21, A_8, A_2, A_1, A_12, A_17, A_23, A_15, A_13, A_14, A_10, A_4, A_18, A_5, A_16, A_7, A_3	76.60
18	A_13, A_3, A_18, A_8, A_21, A_11, A_7, A_15, A_10, A_1, A_4, A_14, A_12, A_17, A_6, A_24, A_2, A_25	80.85
19	A_25, A_8, A_13, A_10, A_22, A_15, A_4, A_17, A_6, A_19, A_3, A_12, A_16, A_14, A_1, A_11, A_20, A_24, A_7	76.60
20	A_19, A_4, A_23, A_16, A_10, A_21, A_14, A_15, A_1, A_2, A_3, A_8, A_11, A_7, A_12, A_18, A_25, A_17, A_13, A_22	76.60
21	A_2, A_6, A_3, A_5, A_14, A_12, A_18, A_10, A_13, A_7, A_22, A_17, A_15, A_8, A_21, A_20, A_1, A_19, A_4, A_11, A_16	76.60
22	A_24, A_1, A_17, A_8, A_20, A_21, A_14, A_12, A_3, A_6, A_19, A_7, A_10, A_13, A_15, A_2, A_23, A_5, A_4, A_11, A_16, A_18	76.60
23	A_25, A_6, A_8, A_15, A_22, A_4, A_12, A_16, A_10, A_17, A_11, A_5, A_1, A_2, A_13, A_21, A_18, A_14, A_3, A_23, A_7, A_19, A_20	76.60

Table 7
Information about datasets.

Dataset name	Number of data	Number of features	Features
DS1	47	25	A_1, …,A_25
DS2	47	5	A_2, A_10, A_11, A_17, A_19
DS3	47	16	A_1, A_2, A_3, A_4, A_5, A_6, A_7, A_8, A_10, A_15, A_16, A_18, A_19, A_22, A_23, A_25

single region. Similarly, this phenomenon has also been emphasized in other psychiatric disorders associated with the 16p11.2 region, such as schizophrenia (Chang et al., 2017, Steinberg et al., 2014). These studies have suggested that both rare and common genetic variants at the 16p11.2 region might increase the risk of schizophrenia.

In the bioinformatic analyses performed in our study, the accuracy rates of dataset 1 consisting of 25 autism and related symptoms, dataset 2 consisting of 5 symptoms and dataset 3 consisting of 16 symptoms were evaluated. The 3 symptoms that we found common in all 3 datasets were as follows: 1. Social and Communication (A2. Delayed response to name), 2. Restricted and Repetitive behaviors (A10. Played with toys or objects in an unusual way (e.g. repetitive play, lining up toys), 3. Aggression (A19. Aggression toward others). Aggression symptoms were remarkable in bioinformatic analysis results similar to bioistatistical analysis results discussed above. In their review of pathogenic CNVs at the 16p11.2 BP4-BP5 region, Natália Oliva-Teles et al. (2020) also examined psychiatric symptoms beyond autism and intellectual disability. Aggression was reported in six studies, with rates ranging from 0 % to 13 %. Among these, Bernier et al. (2017) reported one of the highest rates of aggression associated with 16p11.2, with 30.4 % (7 out of 23 cases).

Recent transcriptomic and proteomic analyses in deletion and duplication models demonstrate that CNV dosage alters key pathways involved in synaptic function and behavioral processes, examining social behavior in two newly developed outbred rat models of the 16p11.2 syndromes associated with ASD (Lorenzo et al., 2023). Integrative gene–phenotype analyses identified Prrt2 as a candidate gene associated with aggressive and stereotyped behaviors. Additional dosage-sensitive genes within the 16p11.2 region, such as Taok2, Kctd13, Sez6l2, and Mapk3, were linked to social isolation. Several genes implicated in hypoactivity and cognitive deficits also showed dosage-dependent dysregulation. These convergent findings support a biologically plausible mechanism whereby CNV-induced perturbation of synaptic and behavioral pathways may contribute to aggression in a subset of carriers.

In addition to the well-established BP4–BP5 breakpoints within the 16p11.2 region that have been widely associated with disease, we also identified recurrent CNVs located more proximally, similar to those observed in our cohort. A search of the DECIPHER database revealed three comparable CNVs reported without additional genomic variants. Phenotypic information was available for

Table 8
The classification performance metrics obtained for the datasets oversampled by SMOTE.

	Accuracy			Precision			Recall			F-measure		
	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3
C4.5	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**
Logistic Regression	73,33	**80**	73,33	84,62	**100**	84,62	**84,62**	76,92	**84,62**	84,62	**86,96**	84,62
Random Forest	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**
Random Tree	80	80	**86,67**	85,71	85,71	**86,67**	92,31	92,31	**100**	88,89	88,89	**92,86**
SVM	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**

Table 9
The classification performance metrics obtained for the datasets oversampled by Borderline SMOTE.

	Accuracy			Precision			Recall			F-measure		
	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3
C4.5	80	80	**86,67**	85,71	85,71	**86,67**	92,31	92,31	**100**	88,89	88,89	**92,86**
Logistic Regression	73,33	**80**	73,33	84,62	**100**	84,62	**84,62**	76,92	**84,62**	84,62	**86,96**	84,62
Random Forest	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**
Random Tree	80	80	**86,67**	85,71	85,71	**86,67**	92,31	92,31	**100**	88,89	88,89	**92,86**
SVM	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**

one of these cases, and it was associated with autism (Table 11). To contextualize our findings, we examined DECIPHER entries overlapping the breakpoints of the common CNVs identified in this study, excluding individuals with multiple CNVs in different genomic regions. Using these criteria, we identified three relevant cases, all summarized in Table 11. Notably, autism spectrum disorder was the only phenotype reported across these DECIPHER entries, suggesting that CNVs located outside the canonical BP4–BP5 interval may also contribute to ASD susceptibility and warrant further attention in clinical interpretation.

One of the major limitations of this study is the lack of detailed clinical information for all cases despite the availability of a large genetic dataset of 7568 individuals. Detailed phenotypic data could be collected only for individuals carrying 16p11.2 CNVs, as restricted by the scope of the ethics approval, and some eligible participants declined further evaluation. Moreover, the small number of pathogenic CNV carriers substantially limits statistical power and the stability of effect-size estimates. Consequently, the observed group differences should be interpreted as preliminary signals rather than definitive genotype–phenotype associations. In line with methodological recommendations for rare-variant research, all significance tests are reported as exploratory, and p-values should be interpreted descriptively rather than confirmatorily. For similar reasons, the machine-learning analyses were conducted as a hypothesis-generating proof-of-concept rather than an attempt to build generalizable predictive models; therefore, classification performance must also be interpreted with caution. Expanding the case series in future studies may enable the formation of more homogeneous genotype groups and thereby improve the robustness of genotype–phenotype correlations.

In conclusion, this study evaluated how both common and rare CNVs within the 16p11.2 region may relate to neurodevelopmental features through detailed clinical assessments of 50 individuals and genomic characterization of 147 CNV carriers. Our findings suggest that structural variation across the broader 16p11.2 locus may contribute to ASD-related phenotypes; however, these results are exploratory and should be interpreted with caution. The evidence presented here highlights a potential widespread contribution of the 16p11.2 interval to behavioral and developmental variability, but larger and systematically phenotyped cohorts will be essential to establish robust, reproducible, and generalizable genotype–phenotype relationships.

Ethic approval statement

Ethics committee approval (Ege University Medical Research Ethics Committee - approval decision 21–9 T/10) was obtained for the study.

Funding Sources

This study was supported by the Scientific and Technological Research Council of Turkiye (TÜBİTAK) under project number

Table 10
The classification performance metrics obtained for the datasets oversampled by ADASYN.

	Accuracy			Precision			Recall			F-measure		
	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3	DS1	DS2	DS3
C4.5	80	80	**86,67**	85,71	85,71	**86,67**	92,31	92,31	**100**	88,89	88,89	**92,86**
Logistic Regression	73,33	**80**	73,33	84,62	**100**	84,62	**84,62**	76,92	**84,62**	84,62	**86,96**	84,62
Random Forest	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**
Random Tree	80	80	**86,67**	85,71	85,71	**86,67**	92,31	92,31	**100**	88,89	88,89	**92,86**
SVM	**86,67**	80	**86,67**	**86,67**	85,71	**86,67**	**100**	92,31	**100**	**92,86**	88,89	**92,86**

Table 11
Clinical Characteristics of DECIPHER-Reported Cases Linking Common CNVS Overlapping Our Breakpoint Region to Autism.

Patient no	Location	Type	Gene	Size	Inheritance	Pathogenity Classification	Phenotype
508976	16:32613257–34158766	Deletion	35	1,55 Mb	Unknown heterozygous	-	No phenotype
293839	16:31974782–33922949	Duplication	37	1.95 Mb	Unknown heterozygous	-	No phenotype
455429	16:31943755–34158766	Duplication	45	2.22 Mb	De novo (unconfirmed parentage) Heterozygous	Uncertain	Autism

Pathogenity Classification was reported in the table according to Decipher descriptions.

321S239.

CRediT authorship contribution statement

Tuba Sözen TÜRK: Software, Resources, Data curation. **Edanur BULUT:** Writing – review & editing, Software, Data curation. **Altuğ KOÇ :** Software, Resources, Data curation. **ÖZGÜL Semiha:** Visualization, Validation, Methodology. **Taha Reşid ÖZDEMIR:** Software, Resources, Data curation. **Duygu Selin TURAN:** Visualization, Validation, Methodology. **Ordin Burak:** Visualization, Validation, Methodology. **Samet ÇELIK:** Project administration, Investigation, Data curation. **Özyılmaz Berk:** Software, Resources, Data curation. **Özgür Ozan KOYUNCU:** Project administration, Investigation, Data curation. **Kosova Buket:** Writing – review & editing, Supervision, Conceptualization. **Özgür KIRBIYIK:** Software, Resources, Data curation. **Özge Özer KAYA:** Software, Resources, Data curation. **Yaşar Bekir KUTBAY:** Software, Resources, Data curation. **Merve Saka GÜVENÇ :** Software, Resources, Data curation. **Kadri Murat ERDOĞAN:** Software, Resources, Data curation. **Bolat Hilmi:** Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **ÜNSEL BOLAT Gül:** Writing – original draft, Project administration, Formal analysis. **Şener ARIKAN:** Software, Resources, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors express their sincere gratitude to the family for their kind participation in this study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.reia.2026.202865.

Data availability

No data was used for the research described in the article.

References

Bijlsma, E. K., Gijsbers, A. C. J., Schuurs-Hoeijmakers, J. H. M., van Haeringen, A., van de Putte, D. F., Anderlid, B. M., & Ruivenkamp, C. A. L. (2009). Extending the phenotype of recurrent rearrangements of 16p11.2. *Deletions in mentally retarded patients without Autism and in Normal individuals European Journal of Medical Genetics, 52*(2–3), 77–87.

Chang, H., Li, L., Li, M., & Xiao, X. (2017). Rare and common variants at 16p11.2 are associated with schizophrenia. *Schizophrenia Research, 184*, 105–108.

Fetit, R., Price, D. J., Lawrie, S. M., & Johnstone, M. (2020). Understanding the clinical manifestations of 16p11.2 deletion syndrome: A series of developmental case reports in children. *Psychiatric Genetics, 30*(5), 136–140.

Medland, S. E., Grasby, K. L., Jahanshad, N., Painter, J. N., Colodro-Conde, L., Bralten, J., Hibar, D. P., Lind, P. A., Pizzagalli, F., Thomopoulos, S. I., Stein, J. L., Franke, B., Martin, N. G., Thompson, P. M., & ENIGMA Genetics Working Group. (2022). Ten years of enhancing neuroimaging genetics through meta-analysis: An overview from the ENIGMA Genetics Working Group. *Human Brain Mapping, 43*(1), 292–299.

Bassuk, A. G., Geraghty, E., Wu, S., Mullen, S. A., Berkovic, S. F., Scheffer, I. E., & Mefford, H. C. (2013). Deletions of 16p11. 2 and 19p13. 2 in a family with intellectual disability and generalized epilepsy. *American Journal of Medical Genetics Part A, 161*(7), 1722–1725.

Mitchell, K. J. (2011). The genetics of neurodevelopmental disease. *Current Opinion in Neurobiology, 21*(1), 197–203.

Niarchou, M., Chawner, S. J. R. A., Doherty, J. L., Maillard, A. M., Jacquemont, S., Chung, W. K., Green-Snyder, L., Bernier, R. A., Goin-Kochel, R. P., Hanson, E., Linden, D. E. J., Linden, S. C., Raymond, F. L., Skuse, D., Hall, J., Owen, M. J., & van den Bree, M. B. M. (2019). Psychiatric disorders in children with 16p11.2 deletion and duplication. *Translational Psychiatry, 9*(1). Article 8.

Oliva-Teles, N., de Stefano, M. C., Gallagher, L., Rakic, S., Jorge, P., Cuturilo, G., Markovska-Simoska, S., Borg, I., Wolstencroft, J., Tümer, Z., Harwood, A. J., Kodra, Y., & Skuse, D. (2020). Rare pathogenic copy number variation in the 16p11.2 (BP4–BP5) region associated with neurodevelopmental and neuropsychiatric disorders: A review of the literature. *International Journal of Environmental Research and Public Health, 17*(24), 9253.

Sicherman, N., Charite, J., Eyal, G., Janecka, M., Loewenstein, G., Law, K., Lipkin, P. H., Marvin, A. R., & Buxbaum, J. D. (2021). Clinical signs associated with earlier diagnosis of children with autism spectrum disorder. *BMC Pediatrics, 21*(1), 1–14.

Steinberg, S., De Jong, S., Mattheisen, M., Costas, J., Demontis, D., Jamain, S., Pietiläinen, O. P. H., Lin, K., Papiol, S., Huttenlocher, J., et al. (2014). Common variant at 16p11.2 conferring risk of psychosis. *Molecular Psychiatry, 19*, 108–114.

Weiner, D. J., Ling, E., Erdin, S., Tai, D. J. C., Yadav, R., Grove, J., Fu, J. M., Nadig, A., Carey, C. E., Baya, N., Bybjerg-Grauholm, J., Berretta, S., Macosko, E. Z., Sebat, J., O'Connor, L. J., Hougaard, D. M., Børglum, A. D., Talkowski, M. E., McCarroll, S. A., Robinson, E. B., & iPSYCH Consortium, ASD Working Group of the Psychiatric Genomics Consortium, & ADHD Working Group of the Psychiatric Genomics Consortium. (2022). Statistical and functional convergence of common and rare genetic influences on autism at chromosome 16p. *Nature Genetics, 54*(11), 1630–1639.

Jolliffe, I. (2011). Principal component analysis. *In International encyclopedia of statistical science* (pp. 1094–1096). Berlin, Heidelberg: Springer.

MacQueen, J. (1967). Multivariate observations. *In Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability, 1*, 281–297.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning (August). *In International conference on intelligent computing* (pp. 878–887). Berlin, Heidelberg: Springer Berlin Heidelberg (August).

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on computational Intelligence)*, 1322–1328 (Ieee).

Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Hosmer, D. W., Jr, Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297.