Liu, Y., & Wei, J.(2023). A systematic review of data mining studies in parenting research. *Journal of Educational Technology Development and Exchange*, *16*(2), 126-144. https://doi.org/10.18785/jetde.1602.08

A Systematic Review of Data Mining Studies in Parenting Research

Yanyan Liu
Tsinghua University, Beijing
yy-liu22@mails.tsinghua.edu.cn

Jun Wei*
Tsinghua University, Beijing
weijun8781@tsinghua.edu.cn

Abstract: *This study aims to provide a comprehensive review of the application of data mining methods in parenting research and to explore the insights offered by data mining algorithms in this domain. In this study, 31 academic parenting research studies, which utilized data mining techniques based on empirical data, are finally selected and thoroughly. They were subsequently analysed, in terms of research inquiries, sources of datasets, and methodological approaches, as well as systematically categorized. Consequently, seven primary research directions in the field of parenting research using data mining methods were outlined. This review also delves into the sources of datasets in extant studies, and discussed the added value that data mining methods bring to the field of parenting research, pointing towards potential directions for future research.*

Keywords: parenting, data mining, big data, machine learning

1. Introduction

Parenting plays a critical role in children's development, as parental influence and guidance have profound and enduring effects on a child's lifelong development. There is a rich history of both theoretical and empirical research in the field of parenting, which has to a certain extent heightened social recognition of the pivotal role of parenting. Researchers have dissected and classified parenting styles (Baumrind, 1967), explored the relationships between parenting styles and children's cognitive, behavioral, and emotional development (Baumrind, 1966; Pomerantz & Wang, 2009; Pomerantz et al., 2014), and analyzed the underlying reasons for parental caregiving behaviors (Florrie & Wei, 2020). These studies have employed traditional quantitative analysis methods such as regression analysis and structural equation modeling to comprehensively investigate the links between parental caregiving and child development. They have provided a theoretical and empirical foundation to guide family education, serving as a cornerstone of significance.

However, due to methodological constraints, these studies have primarily focused on explaining phenomena, with limited research aimed at predicting and preventing children's development problems based on parenting indicators. Furthermore, researchers have yet to utilize the vast textual and non-textual data available on social media and the internet to analyze parenting issues, which places limitations on research in this field.

To address the limitations of traditional research methods and gain a deeper understanding of parenting issues, data mining techniques can be employed. Data mining involves analyzing large datasets to uncover significant patterns, trends, and relationships that may otherwise go unnoticed. The primary objective of data mining is to reveal hidden patterns within data and provide solutions to problems (Wang & Jiang, 2004; Witten & Frank, 2005). Presently, data mining techniques have garnered substantial utilization within the educational domain, thereby germinating a novel interdisciplinary field recognized as educational data mining. Educational data mining is the intersection of computer science, education, and statistics, seeking to apply techniques and methods from statistics, machine learning, and data mining to analyze student learning data and inform educational practices (Lu, 2013).

Educational data mining primarily employs four categories of technical methods, which are predictive algorithms, clustering algorithms, association mining, and topic analysis algorithms. (1) Predictive algorithms are used to estimate the target variable based on existing data. There are several common algorithms in this category, including decision trees, regression, and classification trees. These algorithms can integrate multiple predictor variables to forecast the target variable and help identify which features are crucial in influencing it. (2) Clustering algorithms that are frequently used algorithms here include k-means and hierarchical clustering. Clustering algorithms segment an entire dataset into distinct subsets based on data characteristics, grouping similar data into the same class and dissimilar data into different classes. For example, these algorithms can be used to cluster schools or students to identify similarities and differences between them (Amershi & Conati, 2006; Beal, Qu, & Lee, 2006). Concretely, this method may be deployed to cluster educational institutions, thereby delving into similarities and disparities amongst them, or alternately, to cluster students, thus facilitating an exploration of

the commonalities and disparities among this demographic (Amershi & Conati, 2006; Beal, Qu, & Lee, 2006). (3) Association mining algorithms include correlation analysis, sequence pattern mining, and network analysis. The goal is to identify causal relationships or correlations between variables. (4) Topic analysis algorithms that are frequently used encompass Latent Dirichlet Allocation (LDA) and Structural Topic Model (STM). Topic analysis is employed to uncover the themes within text data, revealing hidden semantic structures and topic distributions. This method analyzes word frequency distributions in text data, automatically detecting latent topic structures within the data, and thereby mitigating the subjective element associated with human judgment. Simultaneously, it serves to economize both time and resources otherwise expended on manual analysis.

Compared to conventional research methods in education, Educational Data Mining (EDM) offers several advantages. Firstly, it utilizes publicly available databases that are characterized by their ecological validity and ease of access. This results in significantly enhanced research efficiency when compared to traditional research methods that often require substantial investments of time and effort in recruiting subjects and entering data (Baker, 2010). Secondly, educational data mining methods enable researchers to swiftly replicate research findings across different datasets. Once an educational pattern is established within a particular dataset, it can be promptly validated and applied to various other datasets (Baker, 2010). Thirdly, educational data mining empowers researchers to construct intricate predictive models that can forecast future outcomes based on existing data. This represents a departure from conventional research paradigms, which are typically focused on explaining observed phenomena (Belgiu & Drăguţ, 2016). Fourthly, this methodology permits the identification and discovery of hidden patterns, trends, and associative relationships within data without the necessity of prior assumptions or hypotheses (Henry et al., 2005). Fifthly, Educational data mining excels in its capacity to effectively manage high-dimensional data and unravel latent relationships between sub-dimensions (Borsboom et al., 2021).

In summary, the application of EDM methods holds paramount significance in transcending the limitations of traditional research paradigms, providing deeper insights into educational issues. It excels in managing vast, multimodal datasets and empowers researchers to make more accurate predictions based on existing data, thereby exerting a more substantial impact on educational practice.

In recent times, a significant number of parents have actively engaged in online forums to articulate their parenting perspectives, interacted through social media platforms, and participated in large-scale surveys. These datasets have built the foundation for conducting research in the domain of parental parenting using data mining methodologies. As the volume of available data continues to grow, there has been a significant increase in the use of data mining methods for parenting research., leading to the emergence of numerous valuable research questions and findings. Although parenting research has been around for a while, the use of data mining methods in this field is still in its early stages. Both theoretical investigations and practical applications are still in their nascent stages of exploration. The true worth and potential of data mining techniques in the context of parental parenting are yet to be fully unearthed, underscoring the pressing need for a greater volume of high-quality research to underpin the future application of data mining

methodologies in this domain. Therefore, it is essential to systematically review and evaluate previous data mining research in the parenting domain, summarize and analyze the major advancements in recent years, dissect the limitations and challenges of existing studies, and delineate future research directions and development paths.

This paper seeks to conduct a comprehensive review of research within the domain of family education that leverages data mining methodologies. It will compile and analyze these studies' data sources, research objectives, and methodological classifications. The article will also reflect on the issues and limitations of existing research in this area. Additionally, it will explore the potential of applying data mining techniques in parenting research, providing a valuable reference for future studies.

2. Methods

2.1. Search for Studies and Inclusion Criteria

To compile articles for this review, we used a two-pronged approach. Firstly, we conducted a comprehensive search of two prominent academic databases: ICNKI and Google Scholar. We restricted our search to articles published before September 1, 2023. This comprehensive search employed two sets of keywords: (1) Terms related to parenting, encompassing aspects such as Parenting, Parents, Fathers, and Mothers; (2) Terms associated with data analysis techniques including Data Mining, Big Data, Random Forest, Latent Dirichlet Allocation (LDA), Structural Topic Model (STM), Text mining, Cluster analysis, Classification, and Decision Trees. Furthermore, we considered studies cited in previously reviewed articles as supplementary sources. Through these combined methods, we initially chose a total of 67 articles.

2.2. Inclusion and Exclusion Criteria

Upon conducting our search, we screened 67 records to ascertain the eligibility of studies for inclusion in our review. These articles were included in our review based on the following criteria : (1)Articles were published in journals. (2) Full-text articles published in international conferences. (3)Present quantitative results.

Conversely, articles were excluded if they met any of the following criteria: (1) Articles that did not rely on empirical data, such as theoretical or conceptual articles, tool demonstrations, and similar types of works. (2) Papers from seminars that exclusively offered abstracts. (3) Book chapters. Following this rigorous screening process, a total of 31 articles remained for our review.

3. Results

This paper analyzes existing literature from three key perspectives: research themes, research methodologies, and research data sources (see Table 1). Among the 31 selected articles, 10 studies have focused on utilizing existing data to predict variables related to children's cognition, behavior, and parental stress. Another 10 studies have analyzed primary parental concerns using textual data, while 6 studies have classified parents based on their parenting styles or behaviors. Furthermore, 5 studies have explored network relationships among various dimensions of parenting. To accommodate the diverse research themes, a wide array of data mining algorithms has been employed in the research methodologies. Specifically, 10 studies have utilized predictive algorithms, including random forests, decision trees, support vector machines, K-nearest neighbors (KNN), XG-BOOST, logistic regression, multiple

regression, and naive Bayes. Another 10 studies have harnessed text mining algorithms, such as Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM). In addition, 6 studies have employed clustering algorithms, including hierarchical clustering and K-means clustering. Furthermore, 5 studies have utilized network analysis methods. Concerning data collection methods, among the 31 studies, 19 have collected data through closed-ended questionnaires, 8 have gathered textual data via web scraping, 2 articles have acquired textual data through open surveys or interviews, and 2 studies have integrated multiple data sources, including open-ended questionnaires, publicly available datasets, and medical diagnoses, for their data analysis endeavors.

Table 1

Summary of Literature on Data Mining in Parenting

Researcher	Research Objective	Research Data	Research Methods
Jayaprakash et al. (2020)	Predicting student academic performance with parenting factors	Closed-ended questionnaire	Naive Bayes, Bagging, Boosting, Random Forest
Saa et al. (2016)	Predicting student academic performance with parenting factors	Closed-ended questionnaire	Decision Trees, Naive Bayes
Sun et al. (2020)	Predicting student academic performance with parenting factors	Closed-ended questionnaire	Logistic Regression, Random Forest
Akter et al. (2023)	Anticipating a child's attachment style and mental health based on parenting styles.	Closed-ended questionnaire	Support Vector Machine, KNN, Random Forest, XG-BOOST, Decision Trees
Halim et al. (2023)	Forecasting the risk of child autism spectrum disorder using parenting factors.	Closed-ended questionnaire	k-means, k-medoid, k-NN, SVM, Artificial Neural Network
Li et al. (2023)	Predicting children's social anxiety using their parents' socioeconomic status.	Closed-ended questionnaire	Bivariate correlation, Logistic Regression, Random Forest
Păsărelu et al. (2022)	Examining the correlation between parental stress and the mental well-being of students.	Closed-ended questionnaire	Network analysis
Florean et al. (2022)	Investigating the links between parenting behaviors and their influence on the mental health of adolescents.	Closed-ended questionnaire	Network analysis
Chen et al. (2010)	Predicting parental stress based on children's developmental challenges and family background.	Closed-ended questionnaire	Decision Trees

Thota et al. (2022)	Examining the correlation between comprehensive parenting styles and child delinquency.	Indian public database	EM clustering algorithm
Calders et al. (2020)	Categorize parenting styles according to psychological, behavioral, and other dimensions.	Closed-ended questionnaire	Subspace k-means clustering
Szymańska & Aranowska (2019)	Categorize parents based on parenting stress, parental expectations, and other relevant variables.	Open-ended questionnaire	Text mining, K-means clustering
Crutzen et al. (2015)	Forecasting child alcohol abuse using parenting styles.	Closed-ended questionnaire	Classification Trees
Liu et al. (2022)	Analysing parenting styles of parents with children at different stages of adolescence and explored the differences between them.	Closed-ended questionnaire	Network analysis
Skjerdingstad et al. (2021)	Analysing the relationship between parental depression symptoms and parenting stress	Closed-ended questionnaire	Network analysis
Blanchard et al. (2021)	Analysing the relationship between parental burnout, marital relationships, and parenting behaviours	Closed-ended questionnaire	Network analysis
Mikkonen et al. (2022)	Categorize Finnish parenting styles based on parental warmth, behavioural control, and psychological control.	Closed-ended questionnaire	Cluster analysis
Martin et al. (2007)	Categorize parenting styles and investigate the connection between parenting styles and child cognition.	Video data coding	Cluster analysis
Park et al. (2021)	Analysing the caregiving requirements of parents of preterm babies in South Korea.	Online text data	Web scraping, LDA
Park et al. (2023)	Analysing parenting questions posed by parents on social media to understand their parenting challenges.	Online text data	Web scraping, LDA
Chung et al. (2022)	Identify the primary parenting strategies adopted by parents during the pandemic, by analysing parents' posts on social media.	Online text data	STM

Lu et al. (2021)	By analysing the text comments on a parenting forum, an analysis was conducted on the school selection information that parents are concerned about.	Online text data	Web scraping, LDA
Li et al. (2021)	Conducting an analysis of the public comments on official media posts on Sina Weibo to identify the concerns of parents.	Online text data	Web scraping, LDA
Westrupp et al. (2022)	Conducting an analysis of parenting inquiries posted by parents on social media to comprehend their parenting difficulties.	Online text data	Web scraping, LDA
Gao et al. (2021)	Analysing parenting questions posed by parents on social media to understand their parenting challenges.	Online text data	Topic analysis, Logistic Regression, BERT
Nelson et al. (2011)	Classify parenting styles with parental control and parental respo nse	Closed-ended questionnaire	K-means clustering
Jigyel et al. (2018)	Investigating Bhutanese parents' involvement in supporting their children with special educational needs (SEN) in schooling.	Interview transcripts	Leximancer topic analysis
He et al. (2020)	Uncovering themes related to parenting needs and questions, as well as discerning the emotional states of parents concealed within their inquiry texts.	Online text data	Web scraping, web text mining
Theule et al. (2011)	Predicting parental stress using children's ADHD symptoms, parental education, social support, and marital status.	Closed-ended questionnaire	Multiple regression
Eli et al. (2021)	Analysing core elements of negative emotions experienced by parents who had lost their only child.	Closed-ended questionnaire	Network analysis
Georg et al. (2021)	Predicting maternal stress with infants' behaviour and family background	Closed-ended questionnaire	Feature selection, GBM

3.1. Data Sources Utilized in Data Mining for Parenting Research

In the field of parenting data mining research, scholars employ various data sources to analyze issues related to parenting. Due to the private nature of family education, a significant number of studies still rely on self-report questionnaires for data. Most researchers opt for closed-ended questionnaires, which require parents or children to respond to surveys regarding family parenting styles, emotional states, adolescent psychological issues, self-concept, and similar aspects (Akter et al., 2023; Calders et al., 2020; Liu et al., 2022; Li et al., 2023; Hsu & Kao, 2007; Skjerdingstad et al., 2021; Blanchard et al., 2021; Buzohre Eli et al., 2021). These questionnaires typically employ rating scales ranging from 1 to 5 or 1 to 7 to precisely measure the emotional and behavioral states of the participants.

While closed-ended questionnaires offer an efficient and convenient means of collecting users' opinions and are more amenable to batch processing, they are designed based on dimensions predetermined by researchers, and the answer choices provided are constrained. This limitation restricts participants' ability to freely express themselves, to some extent, resulting in the loss of a substantial amount of information. Consequently, some researchers have opted for open-ended questionnaires to maximize the retention of participants' authentic perspectives. For instance, Szymańska and Aranowska (2021) and Shima and Kamisawa (2015) furnish participants with open-ended questionnaires that permit them to provide textual responses. These responses are then analyzed through text analysis techniques, examining whether users mention relevant keywords. The collected responses are then reconfigured into a data matrix, facilitating further data analysis.

In addition to traditional questionnaire methods, some researchers have employed web crawling techniques to scrape text from websites. For example, Park et al. (2019), Yazdavar (2017), Lu Kelun et al. (2021), Li Juan et al. (2021), Chung et al., He et al. (2020) all used web crawling methods to obtain a substantial amount of textual data from social media. Regarding the textual data, some studies processed it directly through topic analysis, while others, using key indicators from previous questionnaires, designed keyword dictionaries to encode the text. The encoded data were transformed into subject-indicator matrices for subsequent statistical analysis (Karmen et al., 2015).

Furthermore, some researchers have adopted an integrative approach by combining data from various public databases, textual sources, observational data, and more, for model fitting and analysis. For example, Thota et al. (2022) utilized crime record data from a public database in India. Martin (2007) manually encoded observed parental behaviors from visual recording and conducted further data mining on the encoded data. Karma Jigyel et al. (2018) incorporated textual data from interviews. Anna K. Georg et al. (2021) integrated data from multiple sources, including clinical assessments, questionnaire responses, and records from parental infant diaries. The synthesis of data from multiple sources enhances the depth and scope of research, offering a more comprehensive comprehension and profound insights into the realm of parenting.

3.2. Classification of Parenting Research Topics Using Data Mining Methods

3.2.1. Using predictive algorithms to forecast students' academic performance

Data mining methods can be used to

predict students' academic performance based on parental features. Previous research in parenting has extensively explored the relationship between parenting styles and student academic performance. Parental parenting strategies have emerged as pivotal determinants affecting students' scholastic outcomes, with factors such as parental engagement and parental educational aspirations exhibiting a pronounced correlation with academic achievement (Weiser & Riggio, 2010). Previous studies primarily utilized structural equation models, regression analysis, and other traditional quantitative methods to investigate the impact of family education factors on students' academic performance, elucidating the relationship between parental upbringing and student academic performance (Masud, Thurasamy, & Ahmad, 2015). However, these studies were limited to explaining the importance of existing parental upbringing factors and could not predict students' future academic performance based on existing data. However, educational data mining methods allow researchers to use students' family backgrounds, parenting styles, and other factors to predict students' future academic performance. In such studies, researchers mainly employ prediction algorithms such as random forests, decision trees, support vector machines, and logistic regression. Researchers analyzed different algorithms' accuracy, recall rates, and other metrics to determine the most effective prediction methods for varying conditions. Overall, random forests perform the best among all prediction algorithms. For example, Jayaprakash et al.(2020) predicted students' academic performance using factors like students' academic background and family background, comparing four algorithms: naive Bayes, logit boosting, random forests, and boosted random forests. The results showed that the random forests algorithm had the highest accuracy, reaching 93%. Saa et al. (2016) predicted students' academic performance with social and personal factors and further compared algorithms within random forests like C4.5, ID3, CART, CHAID, etc. The results indicated that the CART algorithm had the best accuracy, with an accuracy rate of 40%.

3.2.2. Using predictive algorithms to forecast adolescent mental health

Data mining techniques can be employed to forecast the psychological well-being of children. Existing research has indicated that children raised by authoritative parents are less likely to encounter psychological problems, whereas children raised under authoritarian parenting styles tend to experience more psychological problems (Szkody, Steele & McKinney, 2021). However, previous research has primarily focused on explaining the causes of psychological issues, rather than proactively predicting them and enabling targeted preventive measures. To prevent psychological health issues in students at an earlier stage, data mining methods utilize variables such as parenting styles, parental emotions, and the family's social economy status to predict a child's personality, psychological well-being, and mental health conditions.

Researchers have employed various prediction algorithms, including random forests, Support Vector Machine (SVM), and k-nearest Neighbor (KNN), among others. The findings of these studies suggest that the efficacy of the algorithm is task-dependent, as it exhibits varying levels of performance across different tasks. For instance, Akter et al. (2023) used the way that parents deal with depression, anxiety, stress, and anger to predict a child's emotional control ability. They compared SVM, KNN, random forests, decision trees, and XG-Boost algorithms, and the results revealed that XG-Boost and random forest classifiers are the most effective

algorithms, achieving an accuracy of 87.3%. Halim et al. (2023) used parental and child historical information to predict the probability of a child being diagnosed with autism. They compared three algorithms: k-nearest neighbors (k-NN), support vector machines (SVM), and artificial neural networks (ANN), finding that SVM is the optimal algorithm for predicting autism with an average accuracy rate of 98.34%. Li et al. (2023) used parental socioeconomic attributes to predict childhood social anxiety disorders. They found that combining logistic regression and random forests effectively predicts childhood social anxiety disorders, with a model accuracy rate of 80.5%.

3.2.3. Predicting parental parenting stress using predictive algorithms

Additionally, data mining methods can be applied to predict parental parenting stress. Past research shows a strong link between parental stress levels and children's physical and psychological health. (Martina et al., 2019). However, these studies can only explain the relationship between parental stress and child-related issues and cannot preemptively predict or intervene in parental stress. The data mining method offers a potential avenue for predicting parental parenting stress.

Various researchers have employed different algorithms to forecast parental stress. For instance, Chen et al. (2010) used decision trees to predict parental stress and found that by constructing a model with variables such as child developmental issues, parental occupation, parental age, parental education level, family annual income, number of children, and child gender, they could effectively predict parental caregiving stress with an accuracy rate of over 75%. Georg et al. (2021), on the other hand, utilized feature selection and Gradient Boosting Machines (GBM) to sift through 464 potential influencing variables, identifying 11 significant factors, to build a predictive model for maternal parenting stress. This model exhibited a favorable predictive performance with an RMSE of 21.72. Theule et al. (2011) employed hierarchical multiple regression to predict parental stress using children's ADHD symptoms, parental education, social support, and marital status. Their model could explain 39% of the total variance. While a direct comparison of algorithmic accuracy proves elusive due to the variances in reported metrics across these studies, it remains evident that all three methodologies efficaciously prognosticate parental stress dynamics.

3.2.4. Analyzing the relationship between overall parenting styles and children development using clustering algorithms

Additionally, data mining techniques can be utilized to evaluate the overall impact of different parenting styles on the cognitive and behavioral development of children. Existing parenting research has uncovered how parental emotions, behavioral control, and psychological control can lead to internalizing and externalizing issues in children (Aunola & Nurmi, 2005). The mentioned studies help understand the connection between specific parenting factors and their impact on children's cognitive and behavioral development. However, parenting style is a complex matter, and children's cognition and behavior are influenced by a combination of various parenting factors instead of the individual effects of each factor. Cluster analysis makes it possible to explore the collective impact of overall parental parenting styles on children. In such studies, families are clustered based on parenting features and family background, followed by an examination of the effects of different families on children's cognition and behavior.

While cluster analysis cannot discern the relative importance of specific factors in influencing children's cognition and behavior, it can consider the combined effects of multiple factors rather than focusing on the individual effects of each factor. This is a notable departure from the conventional approach, which often isolates individual factors for examination, without accounting for their collective dynamics. Consequently, cluster analysis holds critical implications for the predictive modeling of children's cognitive and behavioral profiles. For instance, Calders et al. (2020) used k-means clustering to group parental parenting behaviors based on dimensions like psychology and behavior, resulting in two clusters: authoritative parenting and authoritarian parenting. They then analyzed the occurrence of external problems among adolescents under these two parenting styles. The results revealed a significantly lower probability of external problems among adolescents in the authoritative parenting group compared to those in the authoritarian parenting group.

Szymańska and Aranowska (2019), on the other hand, clustered parents based on variables such as perceived parenting stress, parenting distance, parental personality traits, and parents' expectations for their child's personality. The research deployed structured surveys to measure caregiving stress, parental proximity, and parental personality, while textual descriptions of parental expectations for their offspring's personality were collected through open-ended questionnaires. These descriptions were then recoded using text analysis. Ultimately, a combination of parenting stress, parenting distance, parental personality traits, and parents' expectations for their child's personality were used to cluster parents, resulting in the identification of two groups. The first group of parents exhibited high parenting stress and high parenting distance, with an expectation for their children to exercise more self-control. The second group, on the other hand, displayed low parenting stress, and low parenting distance, and emphasized fostering emotional stability in their children.

3.2.5. Studying parenting networks using network analysis systems

Data mining methods can also provide a systematic analysis of the relationships between various dimensions of parental parenting from a network perspective. Previous research has identified multiple segmented dimensions within the parental caregiving process, including parental involvement, co-parenting, parenting behaviors, and parenting stress. It has been proven that each of these dimensions has a significant impact on the psychological well-being and behavior of children. (Baptista, 2018; Camisasca, 2019). However, traditional approaches tend to ignore the fact that parental caregiving forms a comprehensive system in which these segmented dimensions are interconnected and mutually influential. The data mining method offers a means to explore how these variables interact and exert influence.

Network analysis, a common data mining method used in parenting research, regards psychological structures as elements interacting with each other, collectively forming a network system (Schmittmann et al., 2013). This approach allows for the analysis of relationships between high-dimensional data without any a priori assumptions, enabling researchers to investigate and visualize complex associations between variables (Blanchard & Heeren, 2020). Specifically, network analysis can be used to determine the centrality of each parenting style within the overall caregiving network, discern the interconnections between different parenting styles, and reveal the most effective "target"

factors in parenting (McNally, 2016). For instance, Florean et al. (2022) analyzed the inherent connections between parental caregiving behaviors and explored their impact on the psychological well-being of adolescents. Their findings indicated that within the parenting network, behaviors such as "children staying out late, while parents are unaware" and "your parents praising your good performance" held the most central positions and were the most significant and perceptible parental caregiving behaviors. Ineffective parental monitoring emerged as the predominant parenting style affecting the psychological well-being of adolescents.

Furthermore, network analysis can be a valuable tool for exploring the associations between sub-dimensions of parental parenting and other variables. These studies encompass the sub-dimensions of parental parenting and other target variables within a comprehensive network. For instance, Skjerdingstad et al. (2021) applied network analysis techniques to investigate the network relationships between parental depressive symptoms and parenting stress during the COVID-19 pandemic. Their findings revealed that a sense of worthlessness among parents was a central factor in triggering parental depression and parenting stress. Blanchard (2021) analyzed the interrelationships among parental fatigue, marital relationships, and parenting behaviors. This examination identified emotional distance as a central factor contributing to parental emotional fatigue. Buzohre Eli (2021) delved into the core elements of negative emotions experienced by parents who had lost their only child. The study revealed that apathy, shock, irritability/anger, and night-time nightmares were the most central symptoms among parents who had experienced such a loss. In a similar vein, Păsărelu et al. (2022) scrutinized the relationship between parental parenting stress and children's psychological well-being. Their investigation pinpointed parental perceived coping strategies as the most closely connected nodes in the relationship between parental stress and children's psychological well-being issues.

Furthermore, network analysis can facilitate comparisons between different networks, enabling the examination of the overall structure and global strength of networks, as well as the comparison of variable centrality across different networks. For instance, in a study by Sihan Liu et al. (2022), they analyzed parental caregiving styles at different stages of a child's adolescence. The results revealed that the parenting networks differed among the early, middle, and late stages of adolescence. During the middle stage of adolescence, the interrelationships between parental caregiving variables were stronger than in the early and late stages of adolescence. This implies that during this particular stage, every subtle action taken by parents may potentially have a profound impact on the entire parental caregiving network.

3.2.6. Classifying parenting styles using clustering algorithms

Additionally, data mining methods offer the capability to categorize parental caregiving styles based on data without requiring prior theoretical assumptions. Previous research in parenting has categorized parenting styles, such as Baumrind (1966), who classified parenting styles into authoritarian, permissive, authoritative, and neglectful categories based on the dimensions of parental demandingness and responsiveness. However, these classifications were often based on experiential judgment by researchers. The data mining method, on the other hand, permits researchers to perform clustering of parenting styles without any prior theoretical assumptions, using data from multiple dimensions.

Clustering algorithms in this context are typically divided into hierarchical clustering and k-means clustering. Hierarchical clustering generates a hierarchical structure, allowing for the examination of clustering relationships at different levels, while the k-means algorithm divides data points into non-overlapping clusters, offering lower algorithmic complexity and enhanced interpretability. Given that parenting style classifications are not inherently hierarchical, most researchers chose the k-means clustering approach over hierarchical clustering.

For example, Mikkonen et al. (2022) employed the k-means clustering algorithm to categorize parental caregiving styles based on the dimensions of warmth, behavioral control, and psychological control, resulting in the classification of parenting styles into six categories, including authoritative, permissive, psychologically controlling, uninvolved, controlling, and authoritarian. Through inter-group comparisons, they found that authoritarian parents were more prone to experience parenting fatigue. Similarly, Martin et al. (2007) also utilized the k-means clustering algorithm, clustering parents into high support, partial support, uncooperative-negative, and uncooperative-detached parenting clusters based on dimensions such as parental sensitivity, cognitive stimulation, positive attention, negative attention, interference, and detachment. Nelson et al. (2011), on the other hand, utilized the Ward algorithm and squared Euclidean distance to cluster parenting styles into authoritative, uninvolved, controlling, and unstable categories based on dimensions related to control (psychological control, punishment, verbal hostility, permissiveness) and responsiveness (support, warmth, induction, autonomy granting).

3.2.7. Analyzing parents' primary concerns using text data

Furthermore, parenting data mining leverages text data from sources like online forums to investigate the predominant concerns of parents. Parents produce significant written content across various digital platforms, including social media, forums, and online communities, which reflects their primary interests.

Parenting data mining research employs web scraping techniques to collect relevant textual data and uses topic analysis techniques such as Latent Dirichlet Allocation (LDA) and Structural Topic Modeling (STM) to identify the main interests of parents. Topic analysis automatically discovers latent topic structures within the textual data by analyzing word frequency distributions. This approach obviates the subjectivity inherent in manual categorization and significantly streamlines the labor-intensive and costly manual analysis process. For instance, Ji Hyeon Park et al. (2019) conducted a study to discern the support needs of parents with premature infants in South Korea. They collected and analyzed textual content from posts made by these parents on social networking platforms. Their findings revealed that parents predominantly expressed concerns related to four principal domains: discharge preparations, infant care, child growth and development, and child health issues. Similarly, Sunhwa Park et al. (2023), Elizabeth M. Westrupp et al. (2021), and Yujia Gao et al. (2021) analyzed parents' concerns about social media to understand parenting difficulties. Their research identified feeding, infant food, sleep, medical issues, and child development as the most significant concerns parents faced. Chung et al. (2022) studied parenting strategies employed by parents during pandemic lockdowns. Through the analysis of parents' social media posts,

they identified relevant themes related to parenting strategies during the pandemic, such as spouse support, scheduling, managing expectations, and faith. Additionally, Lu et al. (2021) and Li (2021), through analysis of comment text, discovered that Chinese parents had persistent concerns about children's education resources, academic performance, and academic advancement.

4.Conclusion and Discussion

This study conducted a content analysis of experimental literature in the field of parenting data mining and systematically reviewed and assessed the progress and current status of this field. The research discerns a rapid proliferation within the field of data mining methods used in parenting research, with a primary focus on predicting and analyzing children's academic performance, psychological well-being, cognition, or behavior based on family factors. Additionally, there has been a significant emphasis on classifying parenting styles based on various dimensions of parental upbringing. Besides that, parents' concerns are also explored. Research topics have gradually diversified, with progress observed in areas such as educational outcomes, parental concerns, parenting networks, and parental childcare issues. However, it is worth noting that existing research has some limitations and issues. For example, empirical studies may lack diversity and richness in data sources, and there is room for improvement in terms of the diversity and novelty of research perspectives and methods. Most studies still rely on closed-ended questionnaires, and there are limitations in terms of data quantity. Future research should prioritize addressing educational needs, increasing data collection, promoting interdisciplinary collaboration, and further advancing research in the field of family education.

Based on a comprehensive review of the academic literature, this study outlines several prospective directions for the future of research in the domain of parenting data mining: Firstly, it is anticipated that research in this field will transition from an algorithm-centric paradigm to one that is primarily motivated by educational demands. This shift will emphasize the optimization and precision of algorithmic approaches. Currently, the focus of research still revolves around how to combine data mining methods with parenting topics, in some instances, researchers have overly complexified issues that could have been effectively addressed through conventional statistical techniques such as mean comparisons and linear regression. With the proliferation of parenting data and the maturation of data mining technologies, the complexity of research questions in family education has grown significantly in recent years. Thus, research and application in family education big data will transition from being algorithm-driven to aligning with educational needs, facilitating a shift from algorithmic determinism to educational demand-driven paradigms.

Furthermore, research will focus on diversified, large-scale, and high-quality data collection. Current research primarily relies on closed-ended questionnaires, which have relatively limited data and are constrained by closed responses, resulting in the loss of valuable information. It's worth mentioning that in recent times, many researchers have introduced open-ended questionnaires and have simultaneously considered various types of data sources such as web scraping and public databases, diversifying the data collection methods. The progress made in biometric technology, wearable devices, and other emerging technologies has contributed to the diversification and multimodality of data used in parenting research. This has

paved the way for a more comprehensive and robust data foundation, facilitating related studies in the field. Moreover, future research needs to promote interdisciplinary collaboration and technological integration. Interdisciplinary collaboration and technology integration are important trends in parenting-related data mining research. Parenting-related data mining is an emerging field that involves fundamental knowledge from disciplines such as education and psychology, as well as insights from statistics, information science, cognitive neuroscience, and computer science. Interdisciplinary academic research is crucial for progress beyond the field of education. It is imperative to learn research methods from other disciplines and engage in cross-disciplinary collaboration to achieve this goal. Only in this way can a more comprehensive understanding of data information be achieved, enabling the exploration of more complex parenting issues and providing more precise theoretical and data support for family education.

Author Information

Yanyan Liu
yy-liu22@mails.tsinghua.edu.cn

Jun Wei* Corresponding Author
Institute of Education, Tsinghua University
weijun8781@tsinghua.edu.cn

References

Akter, N., Reza, M. T., & Mostakim, M. (2023, June 7-10). *A Machine Learning Driven Analysis to Determine the Correlation Between Parenting Style and Offspring's Behavior*[Paper presentation]. 2023 IEEE World AI IoT Congress (AIIoT), Seattle, WA, USA, 2023. https://doi.org/10.1109/AIIoT58121.2023.10174459

Amershi, S., & Conati, C. (2006). Automatic Recognition of Learner Groups in Exploratory Learning Environments. In M. Ikeda, K.D. Ashley & T. W. Chan(Eds.) , *Lecture Notes in Computer Science*: vol 4053. *Intelligent Tutoring Systems* (pp. 463–472). Springer. https://doi.org/10.1007/11774303_46

Aunola, K., & Nurmi, J. E. (2005). The role of parenting styles in children's problem behavior. *Child development, 76*(6), 1144-1159. https://doi.org/10.1111/j.1467-8624.2005.00840.x-i1

Baker, R. S. J. D. (2010). Data mining for education. *International Encyclopedia of Education, 7*(3), 112-118.

Baptista, J., Sousa, D., Soares, I., & Martins, C. (2018). Fathers' sensitive guidance moderates the association between co-parenting and behavioral regulation in preschoolers. *International Journal of Behavioral Development, 42*(6), 574-580. https://doi.org/10.1177/0165025418761816

Baumrind, D. (1966). Effects of authoritative parental control on child behavior. *Child Development*, 887-907.

Baumrind, D. (1967). Child care practices anteceding three patterns of preschool behavior. *Genetic Psychology Monographs, 75*(1), 43–88.

Beal, C. R., Qu, L., & Lee, H. (2006). Classifying learner engagement through integration of multiple data sources. In A.G. Cohn(Eds.), *AAAI'06: Proceedings of the 21st national conference on*

Artificial intelligence - Volume 1 (PP. 151–156). https://dl.acm.org/doi/10.5555/1597538.1597563

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing, 114*, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Blanchard, M. A., & Heeren, A. (2020). Why we should move from reductionism and embrace a network approach to parental burnout. *New Directions for Child and Adolescent Development,* 2020(174), 159-168. https://doi.org/10.1002/cad.20377

Blanchard, M. A., Roskam, I., Mikolajczak, M., & Heeren, A. (2021). A network approach to parental burnout. *Child Abuse & Neglect, 111*, 104826. https://doi.org/10.1016/j.chiabu.2020.104826

Borsboom, D., Deserno, M.K., Rhemtulla, M., Epskamp, S., Fried, E.I., McNally, R.J., Robinaugh, D.J., Perugini, M., Dalege, J., Costantini, G., & Isvoranu, A.M. (2021). Network analysis of multivariate data in psychological science. *Nature Reviews Methods Primers, 1*(1), p.58. https://doi.org/10.1038/s43586-021-00055-w

Calders, F., Bijttebier, P., Bosmans, G., Ceulemans, E., Colpin, H., Goossens, L., Van Den Noortgate, W., Verschueren, K., & Van Leeuwen, K. (2020). Investigating the interplay between parenting dimensions and styles, and the association with adolescent outcomes. *European Child & Adolescent Psychiatry, 29*(3), 327-342. https://doi.org/10.1007/s00787-019-01349-x

Camisasca, E., Miragoli, S., & Di Blasio, P. (2019). Children's triangulation during inter-parental conflict: Which role for maternal and paternal parenting stress? *Journal of Child and Family Studies, 28,* 1623-1634. https://doi.org/10.1007/s10826-019-01380-1

Chen, H. Y., Hou, T. W., Chuang, C. H., & TBPS Research Group. (2010). Applying data mining to explore the risk factors of parenting stress. *Expert Systems with Applications, 37*(1), 598-601. https://doi.org/10.1016/j.eswa.2009.05.028

Chung, G., Rodriguez, M., Lanier, P., & Gibbs, D. (2022). Text-mining open-ended survey responses using structural topic modeling: a practical demonstration to understand parents' coping methods during the COVID-19 pandemic in Singapore. *Journal of Technology in Human Services, 40*(4), 296-318. https://doi.org/10.1080/15228835.2022.2036301

Crutzen, R., Giabbanelli, P. J., Jander, A., Mercken, L., & de Vries, H. (2015). Identifying binge drinkers based on parenting dimensions and alcohol-specific parenting practices: building classifiers on adolescent-parent paired data. *BMC Public Health, 15*, 747. https://doi.org/10.1186/s12889-015-2092-8

Eli, B., Liang, Y., Chen, Y., Huang, X., & Liu, Z. (2021). Symptom structure of posttraumatic stress disorder after parental bereavement–a network analysis of Chinese parents who have lost their only child. *Journal of Affective Disorders, 295*, 673-680. https://doi.org/10.1016/j.jad.2021.08.123

Florean, I. S., Dobrean, A., & Roman, G. D. (2022). Early adolescents' perceptions of parenting practices and mental health problems: A network approach. *Journal of Family Psychology, 36*(3), 438-448. https://doi.org/10.1037/fam0000919

Gao, Y., Jang, J., & Yang, D. (2021, May 8 - 13). *Understanding the usage of online media for parenting from infancy to preschool at scale* [Paper presentation]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Yokohama, Japan. https://doi.org/10.1145/3411764.3445203

Georg, A. K., Schröder-Pfeifer, P., Cierpka, M., & Taubner, S. (2021). Maternal parenting stress in the face of early regulatory disorders in infancy: A machine learning approach to identify what matters most. *Frontiers in Psychiatry, 12*, 663285. https://doi.org/10.3389/fpsyt.2021.663285

Halim, Z., Khan, G., Shah, B., Naseer, R., Anwar, S., & Shah, A. (2023). On the Utility of Parents' Historical Data to Investigate the Causes of Autism Spectrum Disorder: A Data Mining-Based Framework. *IRBM, 44*(4), 100780. https://doi.org/10.1016/j.irbm.2023.100780

He, H., He, S., & Li, Y. (2020). Contemporary Chinese parents' needs and questions of parenting for young children: a Web text-mining approach. *Information Research: An International Electronic Journal*, 25. https://doi.org/10.47989/irpaper877

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of Family Psychology, 19*(1), 121-132. https://doi.org/10.1037/0893-3200.19.1.121

Hsu, S. F., & Kao, K. C. (2007, September 5-7). *Contextualization of the relationship between parents, children, and internet by data mining* [Paper presentation]. Second International Conference on Innovative Computing, Information and Control, Kumamoto, Japan. https://doi.org/10.1109/ICICIC.2007.226

Jayaprakash, S., Krishnan, S., & Jaiganesh, V. (2020, March 12-14). *Predicting student's academic performance using an improved random forest classifier* [Paper presentation]. 2020 international conference on emerging smart computing and informatics (ESCI), Pune, India. https://doi.org/10.1109/ESCI48226.2020.9167547

Jigyel, K., Miller, J. A., Mavropoulou, S., & Berman, J. (2018). Parental communication and collaboration in schools with special educational needs (SEN) programs in Bhutan. *International Journal of Inclusive Education, 22*(12), 1288-1305. https://doi.org/10.1080/13603116.2018.1426053

Karmen, C., Hsiung, R. C., & Wetter, T. (2015). Screening internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods. *Computer Methods and Programs in Biomedicine, 120*(1), 27-36. https://doi.org/10.1016/j.cmpb.2015.03.008

Li, G., & Jiang, L. (2023, Febuary 3-5). *Random Forest Algorithm-based Modelling and Neural Network Analysis Between Social Anxiety Disorder of Childhood and Parents' Socioeconomic Attributes* [Paper presentation]. 2023 IEEE 6th Eurasian Conference on Educational Innovation (ECEI), Singapore. https://doi.org/10.1109/ECEI57668.2023.10105416

Li, J., Lu, L., Peng, X., Zeng, Y., & Wang, S. (2023). Critical Discourse Analysis of Public Comments on the Government's Guidance on the Transition from Kindergarten to Primary School by the Big Data Text Mining Technology. *Studies in Early Childhood Education,* (03),70-82. http://doi.org/10.13861/j.cnki.sece.2023.03.008

Liu, S., Xu, B., Zhang, D., Tian, Y., & Wu, X. (2022). Core symptoms and symptom relationships of problematic internet use across early, middle, and late adolescence: a network analysis. *Computers in Human Behavior, 128*, 107090. https://doi.org/10.1016/j.chb.2021.107090

Lu, K., Zhang, P., Chen, B., & Huang, X. (2021). What Does the Choice of Elementary Schools Reflect? Based on the Analysis of Beijing Parents' Comments in BBS. *Peking University Education Review, 19*(03),154-177+192. [in Chinese]

Lu, J. (2013). Big Data and the Application of Education. *Journal of Shanghai*

Educational Research, (9), 5-8+22. [in Chinese] http://doi.org/10.16194/j.cnki.31-1059/g4.2013.09.001

Martin, A., Ryan, R. M., & Brooks-Gunn, J. (2007). The joint influence of mother and father parenting on child cognitive outcomes at age 5. *Early Childhood Research Quarterly, 22*(4), 423-439. https://doi.org/10.1016/j.ecresq.2007.07.001

Martin, C. A., Papadopoulos, N., Chellew, T., Rinehart, N. J., & Sciberras, E. (2019). Associations between parenting stress, parent mental health and child sleep problems for children with ADHD and ASD: Systematic review. *Research in Developmental Disabilities, 93*, 103463. https://doi.org/10.1016/j.ridd.2019.103463

Masud, H., Thurasamy, R., & Ahmad, M. S. (2015). Parenting styles and academic achievement of young adolescents: A systematic literature review. *Quality & Quantity, 49*, 2411-2433. https://doi.org/10.1007/s11135-014-0120-x

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy, 86,* 95-104. https://doi.org/10.1016/j.brat.2016.06.006

Mikkonen, K., Veikkola, H. R., Sorkkila, M., & Aunola, K. (2023). Parenting styles of Finnish parents and their associations with parental burnout. *Current Psychology, 42*(25), 21412-21423. https://doi.org/10.1007/s12144-022-03223-7

Nelson, L. J., Padilla-Walker, L. M., Christensen, K. J., Evans, C. A., & Carroll, J. S. (2011). Parenting in emerging adulthood: An examination of parenting clusters and correlates. *Journal of Youth and Adolescence, 40*, 730-743. https://doi.org/10.1007/s10964-010-9584-8

Ng, F. F. Y., & Wei, J. (2020). Delving into the minds of Chinese parents: What beliefs motivate their learning-related practices? *Child Development Perspectives, 14*(1), 61-67. https://doi.org/10.1111/cdep.12358

Park, J. H., Lee, H., & Cho, H. (2021). Analysis of the supportive care needs of the parents of preterm children in South Korea using big data text-mining: Topic modeling. *Child Health Nursing Research, 27*(1), 34. https://doi.org/10.4094/chnr.2021.27.1.34

Park, S., June, K., & Yu, J. (2023). Analysis of parenting informational needs for mothers with infants and toddlers using text-mining. *Children and Youth Services Review, 145*, 106768. https://doi.org/10.1016/j.childyouth.2022.106768

Păsărelu, C. R., Dobrean, A., Florean, I. S., & Predescu, E. (2023). Parental stress and child mental health: a network analysis of Romanian parents. *Current Psychology, 42,* 24275–24287. https://doi.org/10.1007/s12144-022-03520-1

Pomerantz, E. M., & Wang, Q. (2009). The role of parental control in children's development in Western and East Asian countries. *Current Directions in Psychological Science, 18*(5), 285-289. https://doi.org/10.1111/j.1467-8721.2009.01653.x

Pomerantz, E. M., Ng, F. F. Y., Cheung, C. S. S., & Qu, Y. (2014). Raising Happy Children Who Succeed in School: Lessons From China and the United States. *Child Development Perspectives, 8*(2), 71-76. https://doi.org/10.1111/cdep.12063

Saa, A. A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications, 7*(5). https://doi.org/10.14569/IJACSA.2016.070531

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on

psychological phenomena. *New Ideas in Psychology, 31*(1), 43-53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Shima, M., & Kamisawa, E. (2016). Relevance of reason for breastfeeding to breastfeeding intention in mothers one month after childbirth: Analysis using text mining. *Journal of the Tsuruma Health Science Society Kanazawa University, 39*(2), 125-133.

Skjerdingstad, N., Johnson, M. S., Johnson, S. U., Hoffart, A., & Ebrahimi, O. V. (2021). Feelings of worthlessness links depressive symptoms and parental stress: A network analysis during the COVID-19 pandemic. *European Psychiatry, 64*(1), e50. https://doi.org/10.1192/j.eurpsy.2021.2223

Sun, X., Ram, N., & McHale, S. M. (2020). Adolescent family experiences predict young adult educational attainment: A data-based cross-study synthesis with machine learning. *Journal of Child and Family Studies, 29*, 2770-2785. https://doi.org/10.1007/s10826-020-01775-5

Szkody, E., Steele, E. H., & McKinney, C. (2021). Effects of parenting styles on psychological problems by self-esteem and gender differences. *Journal of Family Issues, 42*(9), 1931-1954. https://doi.org/10.1177/0192513X20958445

Szymańska, A., & Aranowska, E. (2019). Parental stress in the relationship with the child and personality traits that parents shape in their children. *Early Child Development and Care, 191*(2), 198-209. https://doi.org/10.1080/03004430.2019.1611569

Theule, J., Wiener, J., Rogers, M. A., & Marton, I. (2011). Predicting parenting stress in families of children with ADHD: Parent and contextual factors. *Journal of Child and Family Studies, 20*, 640-647. https://doi.org/10.1007/s10826-010-9439-7

Thota, L. S., Shravani, Y., Nagasri, A., Changalasetty, S. B., & Basha, S. A. (2022). *Demarcation of Indian Child Crimes Using Data Mining Approach.* [Paper presentation] 2022 International Conference for Advancement in Technology (ICONAT), Goa, India. https://doi.org/10.1109/ICONAT53423.2022.9725829

Wang, G., & Jiang, P. (2004). Systematic review of data mining. *Journal of Tongji University(Natural Science), 32*(2), 246-252. [in Chinese]

Weiser, D. A., & Riggio, H. R. (2010). Family background and academic achievement: Does self-efficacy mediate outcomes? *Social Psychology of Education, 13,* 367-383. https://doi.org/10.1007/s11218-010-9115-1

Westrupp, E. M., Greenwood, C. J., Fuller-Tyszkiewicz, M., Berkowitz, T. S., Hagg, L., & Youssef, G. (2022). Text mining of Reddit posts: Using latent Dirichlet allocation to identify common parenting issues. *Plos One, 17*(2), e0262529. https://doi.org/10.1371/journal.pone.0262529

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining*: *Practical machine learning tools and techniques (4th Edition)*. Morgan Kaufmann.