Wiley
European Journal of Cancer Care
Volume 2025, Article ID 9070051, 12 pages
https://doi.org/10.1155/ecc/9070051

WILEY

Research Article

Predicting Family Intimacy in Cancer Patients Using Interpretable Machine Learning: Emphasizing Resilience and Self-Esteem

Wen Li,[1,2] **Jingcheng Wen**,[3] **Nuo Zhang**,[1] **Ting Li**,[4] **Hongli Li**,[1] **Yawen Zhang**,[1] **Jie Zhang**,[5] **Yuhan Lu**,[1] **Dong Pang**,[6] **and Hong Yang**[1]

[1]*Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Nursing Department, Peking University Cancer Hospital & Institute, Beijing 100142, China*
[2]*Department of Infectious Diseases, Nanjing Drum Tower Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, Jiangsu 210008, China*
[3]*Nursing Department, Beijing Stomatological Hospital, Capital Medical University, Beijing 100050, China*
[4]*China–Japan Friendship Hospital, Beijing 100029, China*
[5]*Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Gastrointestinal Cancer Center, Unit III, Peking University Cancer Hospital & Institute, Beijing, China*
[6]*School of Nursing, Peking University, Beijing 100083, China*

Correspondence should be addressed to Dong Pang; pangdong@bjmu.edu.cn and Hong Yang; hynl117@sina.com

Received 25 June 2025; Revised 6 November 2025; Accepted 28 November 2025

Academic Editor: Hua Yuan

Objective: This study aimed to construct interpretable machine learning models to predict family intimacy in cancer patients and identify the most influential predictors through SHAP-based analysis.
Methods: A total of 259 cancer patients were surveyed. The data cleaning process involved handling missing values, normalizing continuous variables, and applying one-hot encoding to categorical variables. Statistically significant sociodemographic variables (age, marital status, education, and income) and psychosocial attributes (self-esteem and three resilience subdimensions: tenacity, strength, and optimism) were selected using LASSO regression. Four regression models—gradient boosting (GB), random forest (RF), XGBoost (XGB), and decision tree (DT)—were trained and evaluated using R^2, mean-squared error (MSE), and mean absolute percentage error (MAPE).SHapley Additive exPlanations (SHAP) was used to interpret the GB model.
Results: The GB model achieved the best predictive performance ($R^2 = 0.6985$, MSE = 0.2405), followed by XGB ($R^2 = 0.6794$), RF ($R^2 = 0.6653$), and DT ($R^2 = 0.5912$). SHAP analysis revealed that psychological variables—tenacity, strength, and self-esteem—were the most influential predictors, all exerting strong positive effects. Age group and education showed moderate impact, while income, gender, and marital status contributed minimally.
Conclusion: Gradient boosting offers a robust and interpretable framework for predicting family intimacy in cancer patients. Positive psychological resources—especially resilience and self-esteem—outperform traditional demographics as a predictive foundation, highlighting their clinical significance in survivorship care planning.

Keywords: cancer patients; family intimacy; gradient boosting; machine learning; psychosocial predictors; resilience; self-esteem; SHAP

1. Background

Family intimacy—a central element of family functioning—refers to the emotional bond, support, and closeness that family members experience [1]. In the context of chronic illness, particularly cancer patients, family intimacy not only facilitates emotional coping but also contributes to treatment compliance, psychological resilience, and overall quality of life [2, 3]. Patients who perceive strong emotional connections with family members tend to demonstrate higher levels of hope, reduced distress, and better adjustment to cancer-related challenges [4, 5].

However, the cancer experience can simultaneously pose serious threats to family cohesion. The burden of treatment on the body, coupled with financial and caregiving burdens, often disrupts communication patterns and amplifies stress within the household [6, 7]. This is especially pronounced among older or low-income patients, who may face greater vulnerability to isolation or emotional disconnection [8, 9].

To conceptually ground this study, the Family Systems Theory provides a relevant theoretical lens. This theory posits that family members are interdependent, and changes in one member's emotional or health status inevitably influence the broader family dynamic [10]. Within oncology contexts, cancer-related stress can reshape family roles, emotional exchanges, and coping processes, ultimately affecting levels of family intimacy.

Psychosocial factors have emerged as critical predictors of family intimacy in oncology populations [2, 11]. Among them, self-esteem and resilience are particularly influential. Self-esteem [12], a relatively stable psychological trait, underpins one's sense of worth and ability to maintain meaningful relationships, and has been positively correlated with warmth, trust, and interpersonal sensitivity. Resilience [13], conceptualized as the ability to adapt and recover from adversity, is often divided into dimensions such as tenacity, strength, and optimism. Studies have shown that higher resilience levels buffer against depression [14] and enhance patients' capacity to engage in and sustain intimate family interactions [15].

Despite these findings, the interplay between sociodemographic and psychological variables influencing family intimacy remains insufficiently understood, particularly when accounting for interactive and nonlinear effects. Traditional linear regression models are often limited in their ability to handle multicollinearity, interaction effects, and nonlinear relationships—issues frequently encountered in real-world clinical data [16–18]. In contrast, machine learning (ML) techniques are data-driven, flexible modeling approaches capable of capturing complex, high-dimensional patterns without strict distributional assumptions [19, 20].

Among various ML strategies, supervised regression models such as random forest (RF), gradient boosting, XGBoost, and decision trees have demonstrated superior predictive accuracy in clinical and psychological research [20].

Similar advanced ensemble and XGBoost-based predictive approaches have demonstrated effectiveness in identifying high-risk subgroups and improving decision-making in other complex system domains, such as construction safety risk assessments and BIM-based diagnostic modeling [21, 22], further reinforcing the value of ML for early identification and targeted intervention.

However, there remains a significant research gap: the application of interpretable ML approaches to predict psychosocial outcomes—such as family intimacy—in oncology populations is still limited. Existing studies rarely incorporate model interpretability to translate predictive patterns into clinically meaningful insights.

To address this gap, this study leverages SHapley Additive Explanations (SHAP) to elucidate feature-level contributions, allowing transparent interpretation of ML predictions. Prior studies in other applied domains have shown that combining predictive analytics with model-agnostic interpretability enhances transparency and supports actionable, context-specific decision-making [23], highlighting the importance of interpretability in evidence-based practice. Such interpretability is crucial for translating computational results into personalized psychosocial care plans that are actionable for clinicians.

Importantly, the ability to predict family intimacy offers substantial clinical value. In real-world oncology care, healthcare professionals rarely have the time or resources to conduct in-depth psychosocial assessments for every patient [24, 25]. An accurate predictive model could allow clinicians to identify individuals at risk—even in the absence of direct assessment—and prioritize them for early psychosocial support or family-centered interventions [26, 27].

Accordingly, this study aims to:

1. examine the relative contributions of sociodemographic and psychological factors to family intimacy in cancer patients using multiple ML models;
2. compare the predictive performance of four popular algorithms (RF, gradient boosting, XGBoost, and decision tree); and
3. employ SHAP-based model interpretability to identify the most influential predictors and provide clinically relevant explanations.

The ultimate goal is to provide data-driven and theoretically informed insights that guide targeted psychosocial interventions and enhance family-centered oncology care.

2. Methods

2.1. Study Design and Participants. This study adopted a cross-sectional design to explore the predictive value of psychosocial variables on family intimacy and to compare the performance of multiple ML regression models in this context. A total of cancer patients were recruited from a tertiary oncology hospital using convenience sampling, primarily from the gastrointestinal oncology ward. Participants were diagnosed with various gastrointestinal cancers, including esophageal, gastric, liver, pancreatic, and colorectal cancers. While we were able to collect data on disease stages (e.g., early, advanced), treatment stage information (curative vs. palliative) was not collected due to variations in treatment protocols. This limitation will be addressed in future studies. Inclusion criteria were (1) age ≥ 18 years; (2) ability to understand and complete the questionnaires independently; and (3) informed consent provided voluntarily. Patients with (1) severe psychiatric or cognitive disorders or (2) questionnaires with more than 20% missing responses were excluded.

This study was reviewed and approved by the Ethics Committee of Peking University Cancer Hospital (Approval No.: 2023KT17). Participation is completely voluntary. All participants signed an informed consent form and guaranteed that their personal information would be kept confidential and used only for research purposes.

2.2. Sample Size Estimation. Following established guidance for prediction model development [28], sample size estimation considered both model complexity and the precision required for regression coefficients. Specifically, at least 10–15 observations per predictor variable are generally recommended to minimize overfitting and ensure model reliability. With 10 predictors in the final model, the minimum required sample size was estimated to be 100–150 participants. To further account for potential data loss due to missing responses and to maintain adequate statistical power for internal validation, an additional 20% was added, resulting in a target of approximately 180 cases [29]. The final dataset comprised 259 complete cases after data cleaning, exceeding the calculated requirement and thereby supporting robust model estimation, sufficient power, and reliable cross-validation performance.

2.3. Instruments

2.3.1. Demographic and Sociological Situation Questionnaire. The questionnaire designed by the researchers after the literature review and discussion was adopted, including 12 items: age, age group, disease stage, occupation, gender, marital status, income, payment method of medical expenses, primary caregiver, educational level, understanding of the disease, and presence or absence of cancer pain.

2.3.2. Family Intimacy Scale (FES-CV Intimacy Subscale). Family intimacy was measured using the intimacy subscale of the Chinese version of the Family Environment Scale (FES-CV) [30], revised by Shen et al. in 1991. The scale includes nine items scored on a 5-point Likert scale ranging from 5 ("strongly agree") to 1 ("strongly disagree"), with the first three items reverse-coded. The total score ranges from 9 to 45, with higher scores reflecting greater perceived family intimacy. Previous studies have reported Cronbach's α values consistently above 0.90 for this subscale [31, 32]. Cronbach's α for this subscale in this study was 0.94, demonstrating excellent internal consistency. The Chinese version of the FES also showed satisfactory construct validity, with a KMO value of 0.954 and significant Bartlett's test results ($\chi^2 = 1603.801$, $\mathrm{df} = 36$, $p < 0.001$) in this study.

2.3.3. Self-Esteem Scale (SES). Self-esteem was assessed using the Chinese version of the Rosenberg SES [33], which contains 10 items. Each item is rated on a 4-point Likert scale ranging from 1 ("strongly disagree") to 4 ("strongly agree"). Total scores range from 10 to 40, with higher scores indicating greater trait self-esteem. The scale has shown good reliability in Chinese populations, with Cronbach's $\alpha \geq 0.80$ [34]. Cronbach's α for the SES scale in this study was 0.94, indicating strong internal reliability. Furthermore, the factor structure showed strong construct validity, with a KMO value of 0.958 and significant results in Bartlett's test ($\chi^2 = 1732.282$, $\mathrm{df} = 45$, $p < 0.001$) in this study.

2.3.4. Connor–Davidson Resilience Scale (CD-RISC, Chinese Version). Resilience was measured using the Chinese version of the CD-RISC [35], which contains 25 items across three dimensions: tenacity, strength, and optimism. Items are rated on a 5-point Likert scale (0 = "never true" to 4 = "almost always true"), yielding total scores ranging from 0 to 100. Higher scores indicate greater resilience, and scores are categorized as follows: < 60 = poor, 61–69 = moderate, 70–79 = good, ≥ 80 = excellent. Previous studies have reported Cronbach's α values consistently above 0.80 [36, 37]. In this study, Cronbach's α for the tenacity dimension was 0.965, for the strength dimension was 0.945, and for the optimism dimension was 0.906. The Chinese version of the CD-RISC demonstrated good construct validity, as evidenced by a KMO value of 0.965 and significant Bartlett's test results ($\chi^2 = 5538.451$, $\mathrm{df} = 300$, $p < 0.001$) in this study.

2.4. Data Collection Procedures. All investigators received standardized training from the research team and were required to follow a unified standard operating procedure to ensure questionnaire standardization and consistency across data collectors. Eligible participants were approached during hospitalization and were provided with detailed verbal and written explanations of the study purpose and procedures. After obtaining written informed consent, participants completed the paper-based questionnaire independently in a quiet environment, with investigators available to clarify any questions if necessary.

To ensure inter-rater reliability, questionnaires were reviewed on-site for completeness and logical consistency, including internal logic checks (e.g., reverse-coded items or inconsistent responses) to identify potentially invalid or careless answers. Outliers or implausible values identified during this process were examined by the research team and removed only when they could not be verified against source data, following predefined exclusion criteria to maintain data integrity.

Questionnaires showing uniform or repetitive response patterns were excluded, as were those with more than 10% overall missing data or with over 40% missing values within any individual scale section. All valid questionnaires were double-entered independently by two researchers into a secure database, and discrepancies were resolved through comparison to ensure data accuracy.

2.5. Data Analysis

2.5.1. Data Preprocessing on the Model. Data preprocessing was conducted following a standardized workflow to ensure data quality and model reliability. The main steps are summarized below:

1. Predictor Selection and Statistical Screening

   A two-step screening process was used to identify predictive variables. First, the least absolute shrinkage and selection operator (LASSO) regression was applied using the Python scikit-learn package to address multicollinearity and perform automatic variable selection. The optimal regularization parameter (λ) was determined through a 10-fold cross-validation based on the minimum MSE [38]. Second, the Pearson correlation analysis was performed to examine relationships between family intimacy and key psychological variables (self-esteem and resilience).

2. Data Cleaning and Imputation

   Missing values were handled using mean imputation, which provides a straightforward approach for maintaining dataset completeness and avoiding case deletion. This choice was justified given the small proportion of missing data (< 5%) and the continuous nature of most predictors. However, we acknowledge that more robust imputation techniques (e.g., K-nearest neighbors or multiple imputation) could be adopted in future studies to further improve estimation accuracy [39].

3. Outlier Detection and Treatment

   Outliers were defined as data points exceeding ±3 standard deviations from the mean or showing clear inconsistency upon logical inspection. Such cases were examined against the original records and excluded only if confirmed as data entry or measurement errors [40].

4. Data Transformation

   Categorical variables were converted using one-hot encoding, and all numerical variables were standardized (mean = 0, SD = 1) using the StandardScaler method to improve convergence and comparability across predictors.

5. Data Splitting and Cross-Validation

   The dataset was randomly divided into training (80%) and testing (20%) subsets using stratified sampling to preserve the distribution of the dependent variable (family intimacy).

A 5-fold cross-validation was used within the training set for model tuning and validation, ensuring fold consistency and minimizing performance bias.

Note: As this study focused on a regression task involving continuous outcomes, no oversampling or class balancing techniques were applied [41].

2.5.2. ML Modeling Procedure. All modeling procedures were implemented in Python within the PyCharm 2021 environment [42].

Four regression algorithms were compared for predicting family intimacy: RF, gradient boosting regressor (GBR), Extreme Gradient Boosting (XGBoost), and Decision Tree Regressor (DTR). Hyperparameters were optimized using grid search with 5-fold cross-validation. Model performance was evaluated using R^2, MSE, and mean absolute percentage error (MAPE). Model interpretability was assessed via SHAP analysis, and robustness was examined through residual, Q–Q, and prediction–observation plots.

3. Results

3.1. Variable Selection Using LASSO Regression. The score of family intimacy among cancer patients was 28.58 ± 8.20.

To identify key predictors of family intimacy among cancer patients, Figure 1 displays the coefficient path plot, which illustrates the trajectories of standardized regression coefficients as the regularization penalty (lambda) increases. As lambda increased, coefficients of less relevant variables shrank to zero, ensuring model parsimony and reducing the risk of overfitting.

Figure 2 shows the 10-fold cross-validation results, where the MSE reached its minimum at the optimal lambda value ($\log \lambda \approx -4$). This point was selected for the final model to balance predictive accuracy and variable sparsity.

The final model retained 10 nonzero coefficient predictors: gender, age, age group, education level, marital status, and family income. Their respective coefficients are shown in Table 1. Among them, “age group” ($\beta = 0.4375$) had the largest absolute coefficient, suggesting stronger discriminative power, followed by “Gender” ($\beta = 0.1741$) and “Optimism” ($\beta = 0.1688$). These variables were subsequently used as input features for ML model development.

3.2. Predictive Performance of ML Models. In the univariate analysis, the statistically significant indicators and the indicators with significant correlations were taken as independent variables, and family intimacy was taken as the dependent variable. To evaluate the predictive capability of different algorithms in estimating family intimacy, then four regression models—random forest (RF), gradient boosting (GB), XGBoost (XGB), and decision tree (DT)—were constructed. Figure 3 shows the predicted and actual family intimacy values on the test dataset. Among all models, the prediction curves of GB and RF align most closely with the actual family intimacy values, suggesting better fit. (The summary of the model evaluation indicators can be found in Supporting Information 1).

Figure 4 presents scatter plots of predicted versus actual family intimacy for each model. All scatter points cluster near the ideal diagonal line (in red), indicating good overall prediction performance. The GB model exhibits the highest point concentration around the diagonal, reflecting superior predictive precision.

The residual histograms (Figure 5) show that the residuals of all models approximately follow a normal distribution. The GB model demonstrates the highest kurtosis with residuals more concentrated around zero, implying higher accuracy. Figure 6 presents residual boxplots, where GB and XGB show median values close to zero with fewer outliers, suggesting better model robustness and lower variance.

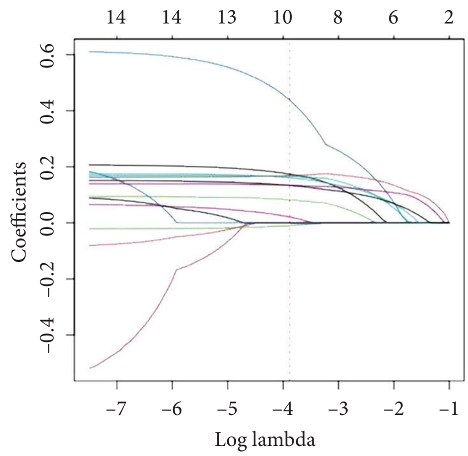

Figure 1: Coefficient path plot.

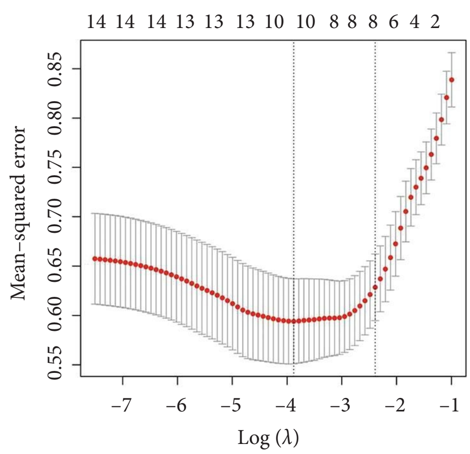

Figure 2: Cross-validation curve.

Table 1: Predictors of family intimacy selected by LASSO regression.

Nonzero coefficient variable	
Variable name	**Coefficient**
Gender	0.17410790
Age	0.01012923
Age group	0.43745522
Education level	0.16514127
Married	0.02092318
Family Income	0.08213909
Self-esteem	0.16016174
Tenacity	0.13402763
Strength	0.13360108
Optimism	0.16883456

In Figure 7, prediction intervals for each model are visualized, highlighting the overlap between predicted and actual values. The GB model displays more stable and narrow confidence intervals (±0.9551), further confirming its high reliability. By contrast, the DT model shows wider intervals (±1.1190), indicating lower prediction certainty.

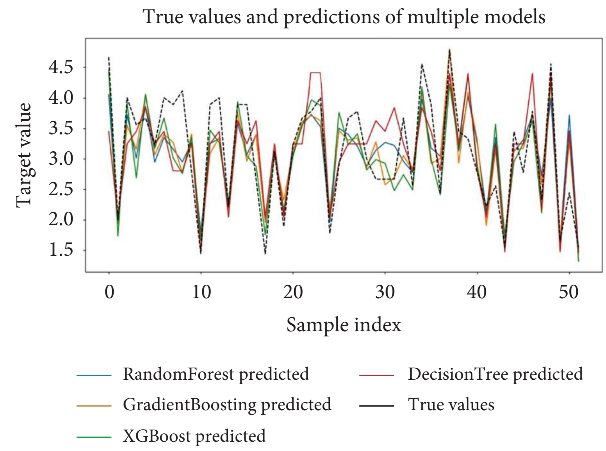

Figure 3: Comparison of the prediction results of multiple models with the true values.

Figure 8 illustrates the residuals plotted against predicted values. All four models show random residual patterns without apparent heteroscedasticity, satisfying regression assumptions. In Figure 9, Q–Q plots of residuals confirm normality, particularly for GB and XGB, whose points adhere closely to the theoretical diagonal.

Aggregated model performance metrics are summarized in Supporting Information 2. GB achieved the highest R^2 value (0.6985), lowest MSE = 0.2405, and lowest MAPE = 0.1320 outperforming RF ($R^2 = 0.6653$, MAPE = 0.1424), XGB ($R^2 = 0.6794$, MAPE = 0.1270), and DT ($R^2 = 0.5912$, MAPE = 0.1454). These results consistently demonstrate that the GB model offers the best overall performance, combining high predictive accuracy with model stability.

3.3. Feature Importance Analysis. We employed SHAP analysis based on the optimal GB model. As illustrated in Figure 10, the features with the greatest impact on family intimacy prediction were psychological resilience-related dimensions and self-esteem. Specifically, tenacity, strength, and self-esteem exhibited the highest SHAP value distributions and exerted consistent positive effects on the model output. Higher values in these attributes were associated with higher predicted levels of family intimacy. These findings underscore the importance of internal psychological strengths in maintaining and fostering close family relationships.

The average absolute SHAP value rankings (Figure 11) provide a clearer picture of relative feature contributions. The top three predictors—tenacity, strength, and self-esteem—demonstrated substantially higher importance scores than other variables, forming the core predictor cluster. These were followed by age group and actual age, suggesting that both developmental stage and chronological age contribute moderately to model predictions. Education level and optimism ranked next, indicating that cognitive resources and positive affectivity play auxiliary roles. At the lower end of the spectrum were gender, family income, and

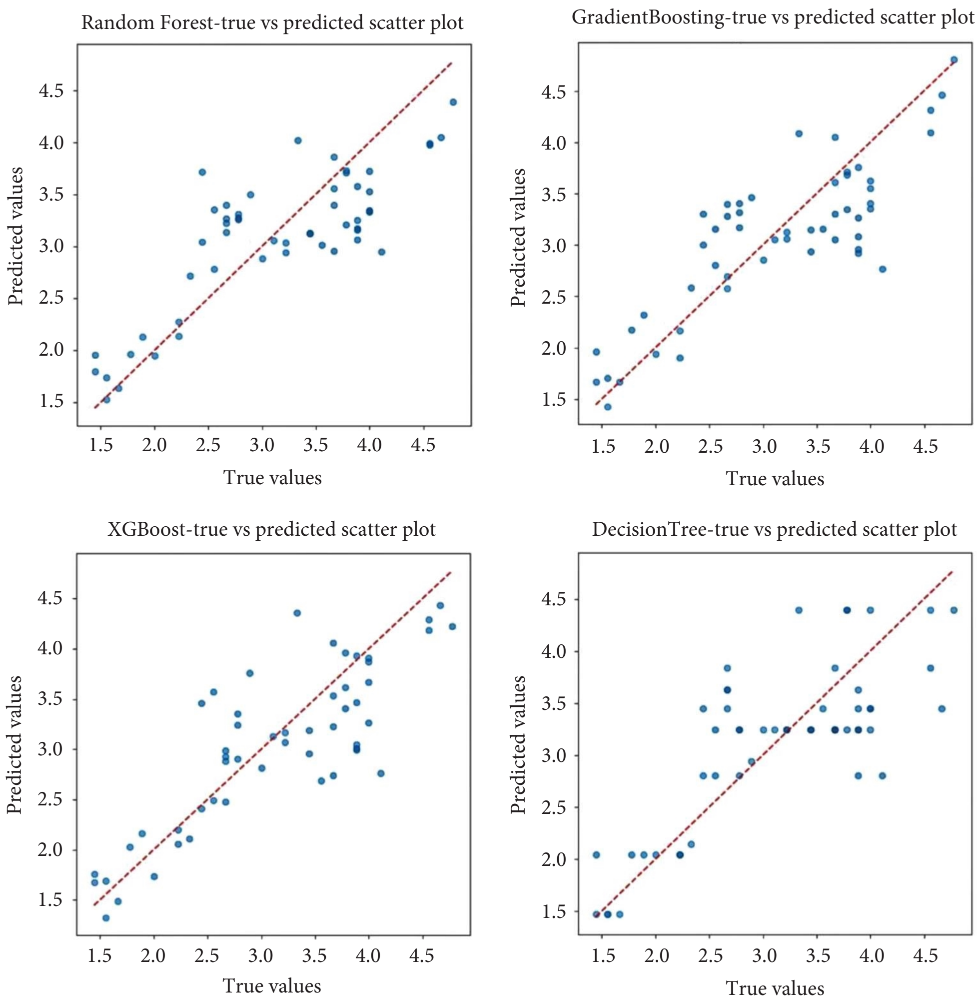

Figure 4: Scatter plots of predicted values versus actual values.

marital status, which showed minimal predictive power in the current dataset.

This gradient of feature importance reflects a broader trend: psychological capital variables outweighed traditional sociodemographic indicators in determining family intimacy.

(Supporting Information 3 shows SHAP dependence plots for top features to explore interaction effects.)

4. Discussion

4.1. Summary of Findings and Model Performance. This study provides an overview of how interpretable ML can identify cancer patients at risk of low family intimacy. Psychological strengths—particularly self-esteem and resilience—emerged as the most influential predictors, outweighing demographic factors. Among the four tested algorithms, GB demonstrated the best performance, with XGBoost following closely, confirming the reliability of ensemble-based models for psychosocial prediction. Integrating SHAP interpretation enhanced transparency, showing that psychological factors contributed substantially more to the model's predictions than demographic variables. Collectively, these findings demonstrate the promise of explainable ML as a practical screening approach to support early psychosocial intervention in oncology.

4.2. Self-Esteem and Resilience as Dominant Predictors. The prominence of self-esteem and resilience aligns with Family Systems Theory, which posits that intrapersonal strengths influence the adaptive functioning of the entire family system.

Consistent with previous studies, high self-esteem is linked to emotional regulation, openness to support, and lower interpersonal withdrawal, enabling patients to sustain close relationships even under cancer-related stress [43–46]. Resilience—particularly tenacity and strength—facilitates emotional recovery, promotes constructive communication, and reinforces mutual support, findings comparable to previous oncology research [15]. Importantly, our results extend this literature by demonstrating that these psychological resources are not merely outcomes of adjustment but serve as predictive mechanisms of family intimacy, reflecting a feedback process emphasized in systemic family theories [47].

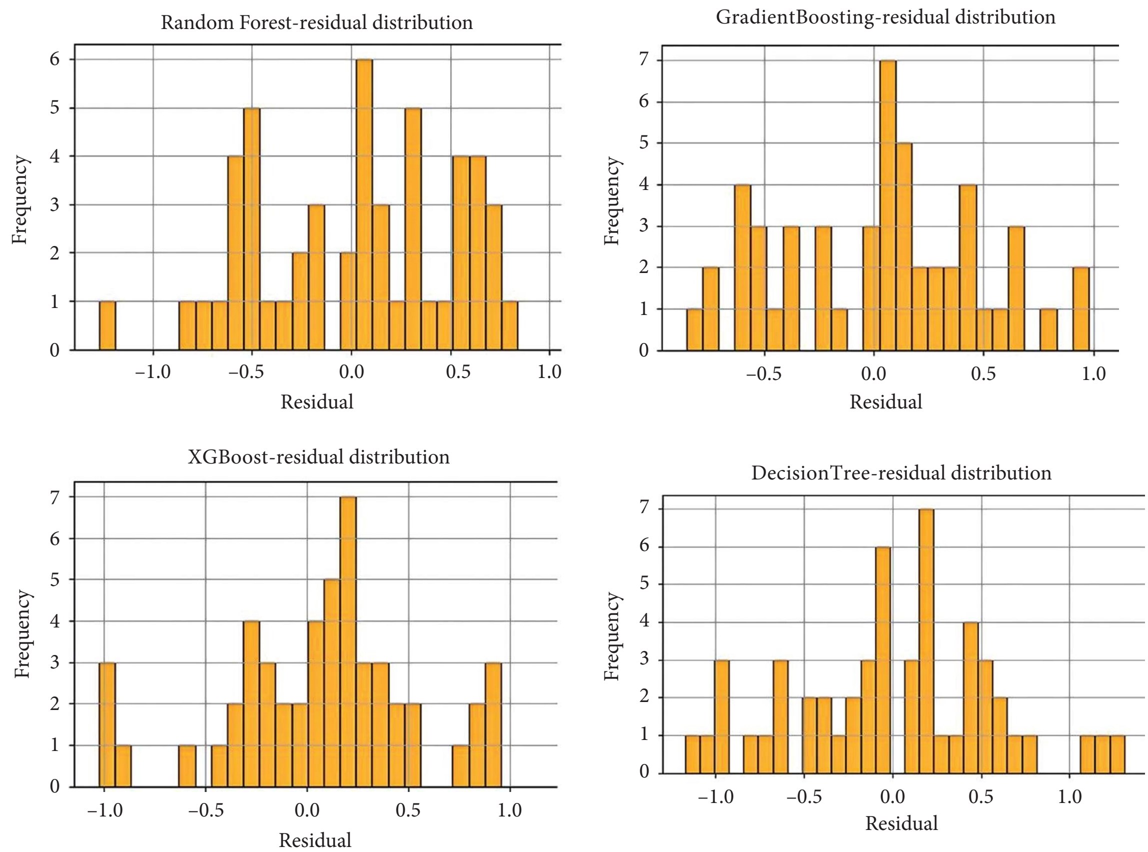

Figure 5: Model residual analysis diagram.

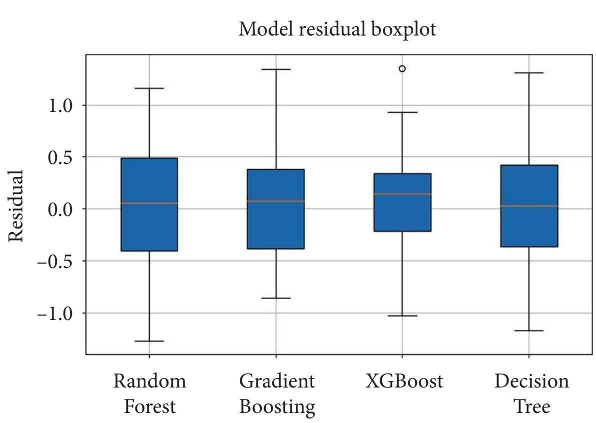

Figure 6: Model residual box plot.

4.3. Demographic Moderators and Sociostructurally Implications. SHAP-based interpretation of the optimal GB model revealed that sociodemographic factors contributed modestly to predictions of family intimacy, with their effects being secondary to psychological variables.

Among the six demographic predictors examined—age, age group, gender, income, marital status, and education level—age-related variables demonstrated relatively higher SHAP values but remained moderate overall. This pattern suggests that developmental stage and life roles may influence relational dynamics only to a limited extent, consistent with prior research emphasizing the transitional nature of intimacy across life phases [48, 49]. Education level emerged as a weak predictor of family intimacy. While education is generally associated with better communication and emotional regulation, its direct effect on familial closeness may be constrained in the context of serious illness, where collective coping rather than individual expression becomes central [50, 51]. This result may also reflect cultural particularities in Chinese family systems. In collectivist cultures, intimacy is often rooted in relational obligations, interdependence, and shared emotional endurance rather than personal achievement or educational attainment [52]. Within this context, family harmony and mutual support are maintained through emotional reciprocity—a process consistent with Family Systems Theory, which emphasizes the dynamic balance between individual adaptation and collective functioning.

Marital status, surprisingly, showed the weakest predictive power, challenging the conventional notion that spousal ties are the foundation of family intimacy [53]. This finding implies that the quality of emotional connection, rather than formal family structure, plays a more critical role in sustaining intimacy when families confront chronic illness. Similarly, gender and income exhibited low SHAP values, indicating a limited explanatory value once psychological resources were accounted for.

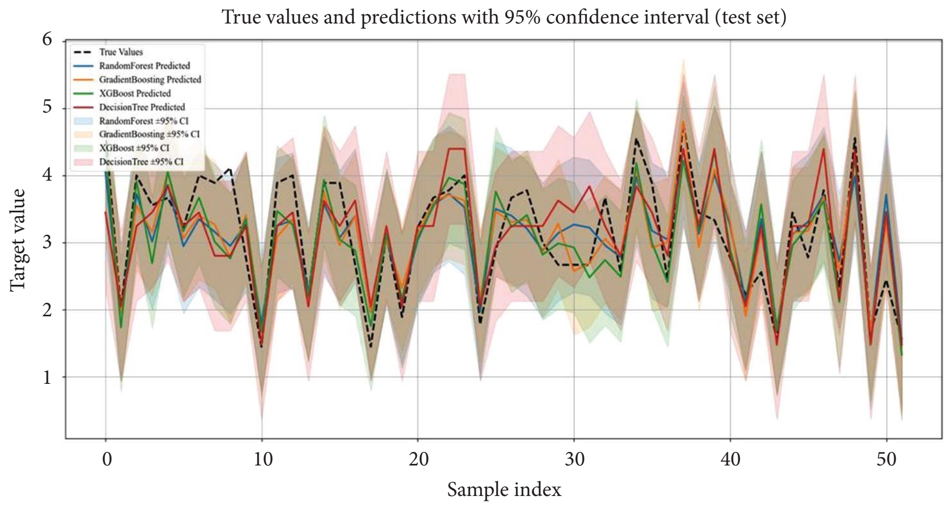

Figure 7: True vs. predicted with 95% CI.

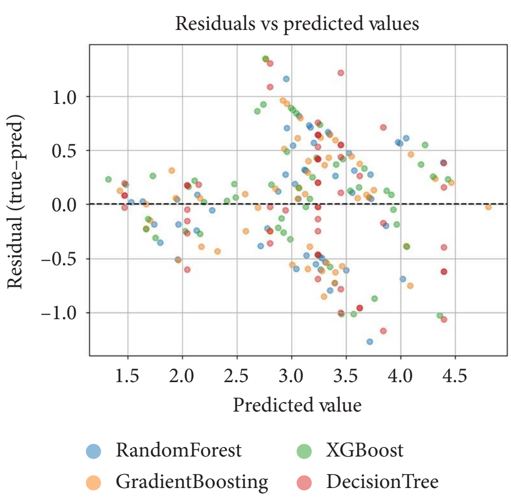

Figure 8: Residuals vs. predicted values.

These results collectively underscore that in Chinese oncology families, internal psychological strengths and relational interdependence are more decisive for maintaining family intimacy than external demographic attributes.

4.4. Methodological Innovations: Explainable AI in Psychosocial Oncology. A notable strength of this study is its use of SHAP-based interpretability to overcome the "black box" limitations of ML [54]. In clinical settings, the lack of transparency in AI-driven predictions has hampered adoption. SHAP enables the assignment of individualized contributions of each variable to each prediction, enabling clinicians to understand not only what is predicted but why [55]. For example, a low family intimacy score can be directly linked to low resilience and high income instability in specific patients, offering immediate clinical insights. While SHAP improved model transparency, it also presents limitations. The additive nature of SHAP values does not capture higher-order feature interactions unless explicitly modeled. Thus, while SHAP enhances interpretability, it should be complemented by interaction analyses or hierarchical modeling to fully understand complex psychosocial relationships.

Nevertheless, its ability to assign feature-level importance and visualize individualized predictions remains a valuable step toward interpretable AI in psycho-oncology [56]. The results of the SHAP underscore the potential of explainable ML to support targeted psychosocial interventions for improving family intimacy in cancer patients.

4.5. Clinical Relevance: Informing Targeted Interventions. The findings of this study provide important implications for psycho-oncological practice by demonstrating how interpretable ML can complement traditional psychosocial assessment. Family intimacy, long recognized as a protective factor for adjustment, treatment adherence, and quality of life [4, 57], can now be evaluated not only through self-report instruments but also via predictive modeling that identifies patients at risk for relational strain. The model's predictive function offers clinicians a screening mechanism for early detection of patients exhibiting psychosocial vulnerability. For example, patients with low predicted intimacy scores driven by low resilience or low self-esteem can be proactively referred for targeted psychological services even before overt distress is reported.

From a practical standpoint, the model could be embedded into routine electronic health records (EHRs) or patient management systems, allowing real-time risk stratification during oncology visits. Such integration would enable multidisciplinary teams—oncologists, nurses, and psychologists—to coordinate timely family-centered interventions without increasing the clinical workload.

Interventions aimed at strengthening self-esteem and resilience—such as cognitive–behavioral therapy, resilience-building workshops, and family-based psychoeducation [58–60]—may be prioritized for those flagged by the model as high risk.

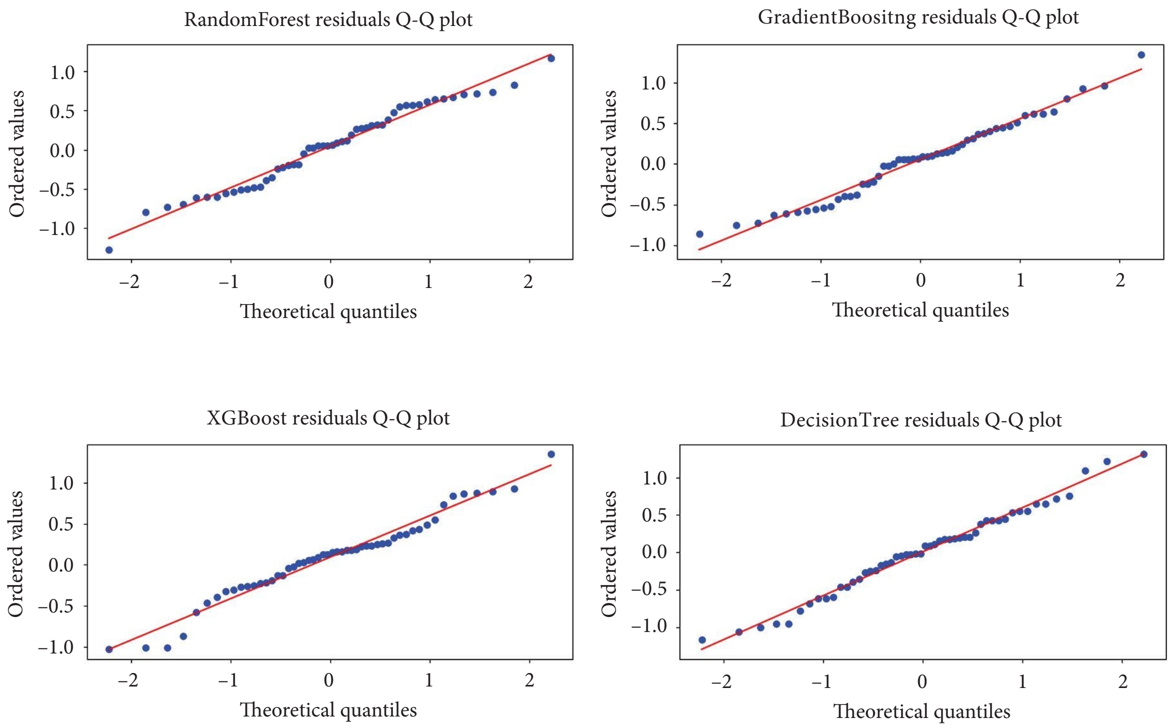

Figure 9: Q–Q plots.

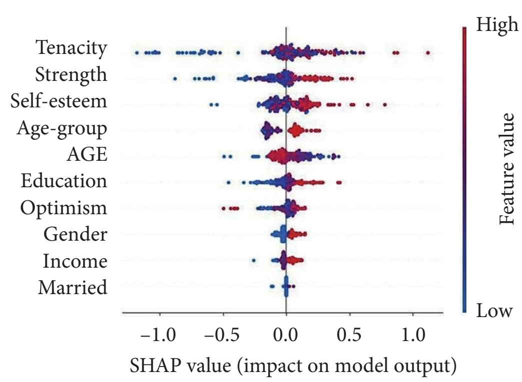

Figure 10: SHAP summary plot.

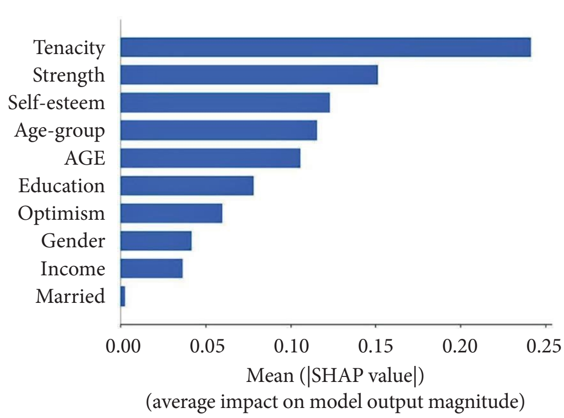

Figure 11: Mean SHAP value bar plot.

Moreover, the SHAP interpretability framework provides individualized insight, helping clinicians explain to patients and families why specific psychosocial dimensions contribute to risk, thereby enhancing communication and patient engagement. Future applications may extend beyond oncology to other chronic disease contexts where family dynamics influence coping and recovery. By integrating predictive analytics with person-centered care, this approach bridges data science and clinical psychology, moving toward a precision psychosocial model of healthcare delivery.

5. Limitations and Future Prospects

This study has several limitations. First, the cross-sectional design limits our ability to infer causal relationships between psychosocial factors and perceived family intimacy. The potential for overfitting still existed despite cross-validation. Inability to capture temporal or dynamic changes in family intimacy. Second, although the sample size met the minimum requirements for ML modeling, it remains relatively modest. This may restrict the generalizability of the findings to broader cancer populations with diverse clinical and demographic characteristics. Because participants were recruited via convenience sampling at a single tertiary center, selection bias and limited representativeness are possible. Third, while SHAP analysis enhances model interpretability, it does not capture potential interaction effects between features unless explicitly modeled. As such, synergistic relationships between psychosocial variables—such as how resilience and self-esteem might jointly influence intimacy—may be underestimated. Lastly, this study focused solely on structured data and lack of unstructured data; incorporating unstructured data sources such as narrative texts or audio interviews may further enrich model performance and contextual insight into future research.

Future studies should involve longitudinal follow-up and consider stratified or quota sampling across cancer types and treatment stages to enhance generalizability. In subsequent applied research, when the ML model be integrated into clinical workflows, EHRs, the ethical risks in clinical deployment of predictive models (bias, data privacy) should also not be overlooked. It is necessary to strengthen clinician training on information systems to promote model acceptance and ensure the smooth advancement of the system. Additionally, further collecting real-world data will optimize the model, thereby better addressing bias issues in predictive models, particularly considering variations brought about by demographic variables.

6. Conclusion

This study applied four interpretable ML models to predict perceived family intimacy among cancer patients. Among the four models tested, the GB algorithm demonstrated the highest predictive accuracy and robustness. SHAP analysis revealed the dominant role of resilience and self-esteem in prediction. These findings highlight the critical role of internal psychological strengths in improving family relationships within the cancer context and provide clinical practice guidelines for targeted screening of low family intimacy and intervention research. Future research should incorporate longitudinal designs, larger and more diverse samples, and multimodal data sources, and conduct applied research to integrate it into clinical practice, so as to optimize and enhance its applicability.

Data Availability Statement

The datasets used during the current study are available from the corresponding author upon reasonable request.

Ethics Statement

This research was approved by the Institutional Review Board of Peking University Cancer Hospital (approval No. 2023KT17). Informed consent was obtained from all participants.

Disclosure

All authors read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Author Contributions

Wen Li, Jingcheng Wen: conceptualization, data analysis, and writing original draft.

Dong Pang and Hong Yang: data analysis method guidance, writing original draft project supervision, and manuscript revision.

Nuo Zhang, Ting Li, Hongli Li, Yawen Zhang, and Jie Zhang: data collection, methodology, and writing—review and editing.

Yuhan Lu: project supervision and manuscript revision.

Wen Li and Jingcheng Wen are co-first authors and have contributed equally to the article.

Funding

This work was supported by the Science Foundation of Peking University Cancer Hospital (PY202336) and the National Public Hospital Reform and High-quality Development Demonstration Project (Haidian District, Beijing) (11010822T000001740154).

Acknowledgments

The authors thank all participants and staff for their support and cooperation during the study.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. *(Supporting Information)*

1: Summary of the model evaluation indicators. 2: Summary of performance indicators of the aggregated model. 3: SHAP dependence plots for top features to explore interaction effects.

References

[1] M. Lanz and S. Tagliabue, *Family Intimacy Measures. Encyclopedia of Quality of Life and Well-Being Research* (Springer, 2024).

[2] S. Manne and H. Badr, "Intimacy and Relationship Processes in Couples' Psychosocial Adaptation to Cancer," *Cancer* 112, no. S11 (2008): 2541–2555, https://doi.org/10.1002/cncr.23450.

[3] E. C. Jiakponna, J. O. Agbomola, O. Ipede, et al., "Psychosocial Factors in Chronic Disease Management: Implications for Health Psychology," *International Journal of Science and Research Archive* 12, no. 2 (2024): 117–128, https://doi.org/10.30574/ijsra.2024.12.2.1219.

[4] C. Li, J. Yuan, X. Huang, S. Zhang, Y. Hong, and J. Zhong, "Correlation Between Depression and Intimacy in Lung Cancer Patients and Their Family Caregivers," *BMC Palliative Care* 21, no. 1 (2022): 99, https://doi.org/10.1186/s12904-022-00992-7.

[5] E. M. Manczak, K. A. Skerrett, L. B. Gabriel, K. A. Ryan, and S. A. Langenecker, "Family Support: A Possible Buffer Against Disruptive Events for Individuals With and Without Remitted Depression," *Journal of Family Psychology* 32, no. 7 (2018): 926–935, https://doi.org/10.1037/fam0000451.

[6] M. Bevans and E. M. Sternberg, "Caregiving Burden, Stress, and Health Effects Among Family Caregivers of Adult Cancer Patients," *JAMA* 307, no. 4 (2012): 398–403, https://doi.org/10.1001/jama.2012.29.

[7] E. M. Johnson, “Stress and Cancer Treatment: How a Family’s Adaptability and Cohesion Affects Financial Stress,” (Lexington, Kentucky: University of Kentucky, 2020), Theses and dissertations.

[8] X. Meng, Z. An, Y. Xu, et al., “Living Experiences of People With Advance Cancer With Low Socioeconomic Status: A Systematic Review of Qualitative Evidence,” *Palliative Medicine* 37, no. 4 (2023): 444–459, https://doi.org/10.1177/02692163221137106.

[9] S. Hasan, *Family Resilience and the Lived Experience of Pancreatic Cancer: A Multiple Case Study Design* (Saint Louis University, 2024).

[10] F. Rothbaum, K. Rosen, T. Ujiie, and N. Uchida, “Family Systems Theory, Attachment Theory, and Culture,” *Family Process* 41, no. 3 (2002): 328–350, https://doi.org/10.1111/j.1545-5300.2002.41305.x.

[11] R. Caruso, M. G. Nanni, M. B. Riba, S. Sabato, and L. Grassi, “The Burden of Psychosocial Morbidity Related to Cancer: Patient and Family Issues,” *International Review of Psychiatry* 29, no. 5 (2017): 389–402, https://doi.org/10.1080/09540261.2017.1288090.

[12] M. A. Harris and U. Orth, “The Link Between Self-Esteem and Social Relationships: A Meta-Analysis of Longitudinal Studies,” *Journal of Personality and Social Psychology* 119, no. 6 (2020): 1459–1477, https://doi.org/10.1037/pspp0000265.

[13] Y. Zhang, Y. Wu, and Y. Li, “Sex Differences in the Mediating Effect of Resilience on Social Support and Cognitive Function in Older Adults,” *Geriatric Nursing* 53 (2023): 50–56, https://doi.org/10.1016/j.gerinurse.2023.06.013.

[14] K. T. Laird, B. Krause, C. Funes, and H. Lavretsky, “Psychobiological Factors of Resilience and Depression in Late Life,” *Translational Psychiatry* 9, no. 1 (2019): 88, https://doi.org/10.1038/s41398-019-0424-7.

[15] A. Seiler and J. Jenewein, “Resilience in Cancer Patients,” *Frontiers in Psychiatry* 10 (2019): 208, https://doi.org/10.3389/fpsyt.2019.00208.

[16] N. C. Chesnaye, M. van Diepen, F. Dekker, C. Zoccali, K. J. Jager, and V. S. Stel, “Non-Linear Relationships in Clinical Research,” *Nephrology Dialysis Transplantation* 40, no. 2 (2025): 244–254, https://doi.org/10.1093/ndt/gfae187.

[17] J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, et al., “Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review,” *Mathematics* 10, no. 8 (2022): 1283, https://doi.org/10.3390/math10081283.

[18] D. Maulud and A. M. Abdulazeez, “A Review on Linear Regression Comprehensive in Machine Learning,” *Journal of applied science and technology trends* 1, no. 2 (2020): 140–147, https://doi.org/10.38094/jastt1457.

[19] A. Wilson and M. R. Anwar, “The Future of Adaptive Machine Learning Algorithms in High-Dimensional Data Processing,” *International Transactions on Artificial Intelligence* 3, no. 1 (2024): 97–107, https://doi.org/10.33050/italic.v3i1.656.

[20] E. Dritsas and M. Trigka, “Efficient Data-Driven Machine Learning Models for Cardiovascular Diseases Risk Prediction,” *Sensors* 23, no. 3 (2023): 1161, https://doi.org/10.3390/s23031161.

[21] A. Shehadeh and O. Alshboul, “Enhancing Occupational Safety in Construction: Predictive Analytics Using Advanced Ensemble Machine Learning Algorithms,” *Engineering Applications of Artificial Intelligence* 159 (2025): 111761, https://doi.org/10.1016/j.engappai.2025.111761.

[22] A. Shehadeh, O. Alshboul, and M. Arar, “Enhancing Urban Sustainability and Resilience: Employing Digital Twin Technologies for Integrated WEFE Nexus Management to Achieve SDGs,” *Sustainability* 16, no. 17 (2024): 7398, https://doi.org/10.3390/su16177398.

[23] A. Shehadeh and O. Alshboul, “Enhancing Engineering and Architectural Design Through Virtual Reality and Machine Learning Integration,” *Buildings* 15, no. 3 (2025): 328, https://doi.org/10.3390/buildings15030328.

[24] C. K. B. Muralidhara, “Interpretability of Classification & Regression Ensemble Models,” (Tampere, Finland: Tampere University, 2024), Master of science thesis.

[25] S. Mangalathu, S.-H. Hwang, and J.-S. Jeon, “Failure Mode and Effects Analysis of RC Members Based on Machine-Learning-Based Shapley Additive exPlanations (SHAP) Approach,” *Engineering Structures* 219 (2020): 110927.

[26] C. G. Ratcliff, C. A. Vinson, K. Milbury, and H. Badr, “Moving Family Interventions Into the Real World: What Matters to Oncology Stakeholders?” *Journal of Psychosocial Oncology* 37, no. 2 (2019): 264–284, https://doi.org/10.1080/07347332.2018.1498426.

[27] F. Davies, M. Harris, H. L. Shepherd, et al., “Promise Unfulfilled: Implementing Web-Based Psychological Therapy in Routine Cancer Care, a Qualitative Study of Oncology Health Professionals’ Attitudes,” *Psycho-Oncology* 31, no. 7 (2022): 1127–1135, https://doi.org/10.1002/pon.5900.

[28] R. D. Riley, K. I. Snell, J. Ensor, et al., “Minimum Sample Size for Developing a Multivariable Prediction Model: Part I–Continuous Outcomes,” *Statistics in Medicine* 38, no. 7 (2019): 1262–1275, https://doi.org/10.1002/sim.7993.

[29] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider, “Missing Data in Surveys: Key Concepts, Approaches, and Applications,” *Research in Social and Administrative Pharmacy* 18, no. 2 (2022): 2308–2316, https://doi.org/10.1016/j.sapharm.2021.03.009.

[30] F. Shen, Y. Zheng, Z. Jingping, S. Jiang, L. Wang, and X. Wang, “Preliminary Evaluation of the ‘Family Intimacy and Adaptability Scale’ and the ‘Family Environment Scale’—A Comparative Study Between Normal Families and Families of Schizophrenic Patients,” *Chinese Journal of Mental Health* no. 5 (1991): 198–202 + 38.

[31] F. Xiaoyan, X. Xiaotian, and M. Zhu, “The Impact of Structured Family Therapy on the Family Structure and Function of Children With Hematological Malignancies,” *Chinese Journal of Contemporary Pediatrics* 16, no. 9 (2014): 905–909.

[32] J. Jiang, Y. Zhong, B. Gao, et al., “Characteristics of Family Environment of Children with Attention Deficit Hyperactivity Disorder and Emotional Instability,” *Chinese Mental Health Journal/Zhongguo Xinli Weisheng Zazhi* 35, no. 4 (2021).

[33] M. Rosenberg, “Rosenberg Self-Esteem Scale (RSE),” *Acceptance and Commitment Therapy/Measures Package* (1965): 61.

[34] B. Curbow and M. Somerfield, “Use of the Rosenberg Self-Esteem Scale With Adult Cancer Patients,” *Journal of Psychosocial Oncology* 9, no. 2 (1991): 113–131, https://doi.org/10.1300/j077v09n02_08.

[35] K. M. Connor and J. R. Davidson, “Development of a New Resilience Scale: The Connor-Davidson Resilience Scale (CD-RISC),” *Depression and Anxiety* 18, no. 2 (2003): 76–82, https://doi.org/10.1002/da.10113.

[36] Ö. Karaırmak, “Establishing the Psychometric Qualities of the Connor–Davidson Resilience Scale (CD-RISC) Using Exploratory and Confirmatory Factor Analysis in a Trauma Survivor Sample,” *Psychiatry Research* 179, no. 3 (2010): 350–356, https://doi.org/10.1016/j.psychres.2009.09.012.

[37] S. Bezdjian, K. G. Schneider, D. Burchett, M. T. Baker, and H. N. Garb, "Resilience in the United States Air Force: Psychometric Properties of the Connor-Davidson Resilience Scale (CD-RISC)," *Psychological Assessment* 29, no. 5 (2017): 479–485, https://doi.org/10.1037/pas0000370.

[38] Y. Huang, M. Cho, S. Chakraborty, and T. Dey, "Variable Selection for Prediction in Clinical Research," *Wiley Interdisciplinary Reviews: Computational Statistics* 17, no. 2 (2025): e70030, https://doi.org/10.1002/wics.70030.

[39] J. Poulos and R. Valle, "Missing Data Imputation for Supervised Learning," *Applied Artificial Intelligence* 32, no. 2 (2018): 186–196, https://doi.org/10.1080/08839514.2018.1448143.

[40] M. M. Ahsan, M. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies* 9, no. 3 (2021): 52, https://doi.org/10.3390/technologies9030052.

[41] D. Rastogi, P. Johri, V. Tiwari, and A. A. Elngar, "Multi-Class Classification of Brain Tumour Magnetic Resonance Images Using Multi-Branch Network With Inception Block and Five-Fold Cross Validation Deep Learning Framework," *Biomedical Signal Processing and Control* 88 (2024): 105602, https://doi.org/10.1016/j.bspc.2023.105602.

[42] S. Kotha, *A Study on the Impact of Preprocessing Steps on Machine Learning Model Fairness* (Purdue University Graduate School, 2024).

[43] B. Fernandes, J. Newton, and C. A. Essau, "The Mediating Effects of Self-Esteem on Anxiety and Emotion Regulation," *Psychological Reports* 125, no. 2 (2022): 787–803, https://doi.org/10.1177/0033294121996991.

[44] N. A. Ruvalcaba-Romero, P. Fernández-Berrocal, J. G. Salazar-Estrada, and J. Gallegos-Guajardo, "Positive Emotions, Self-Esteem, Interpersonal Relationships and Social Support as Mediators Between Emotional Intelligence and Life Satisfaction," *Journal of Behavior, Health & Social Issues* 9, no. 1 (2017): 1–6, https://doi.org/10.1016/j.jbhsi.2017.08.001.

[45] Y. Lei, M. Li, C. Lin, C. Zhang, and Z. Yu, "The Effect of Ostracism on Social Withdrawal Behavior: The Mediating Role of Self-Esteem and the Moderating Role of Rejection Sensitivity," *Frontiers in Psychology* 15 (2024): 1411697, https://doi.org/10.3389/fpsyg.2024.1411697.

[46] L. Sand, "Existential Challenges and Coping in Palliative Cancer Care: Experiences of Patients and Family Members: Karolinska Institutet" (2008).

[47] D. P. Berthold and M. Bormann, *Psychological Resilience Promotes Positive Outcomes After Surgery* (Elsevier, 2024).

[48] G. O. Hagestad, "Interdependent Lives and Relationships in Changing Times: A Life-Course View of Families and Aging," *Invitation to the Life Course* (Routledge, 2018), 135–159.

[49] T. Liu, J. Nicholas, M. M. Theilig, et al., "Machine Learning for Phone-Based Relationship Estimation: the Need to Consider Population Heterogeneity," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, no. 4 (2019): 1–23, https://doi.org/10.1145/3369820.

[50] R. E. Cole and D. Reiss, *How Do Families Cope With Chronic Illness?* (Routledge, 2013).

[51] J. M. Patterson and A. W. Garwick, "The Impact of Chronic Illness on Families: A Family Systems Perspective," *Annals of Behavioral Medicine* 16, no. 2 (1994): 131–142, https://doi.org/10.1093/abm/16.2.131.

[52] N. V. Roman, T. V. Balogun, L. Butler-Kruger, et al., "Strengthening Family Bonds: A Systematic Review of Factors and Interventions That Enhance Family Cohesion," *Social Sciences* 14, no. 6 (2025): 371, https://doi.org/10.3390/socsci14060371.

[53] A. Batool, A. Aziz, N. Shafiq, and T. ur Rahman, "Trust, Intimacy, and Their Impact on Spousal Relationships Among Older Adults: Examining Key Psychological Dynamics," *Social Science Review Archives* 3, no. 1 (2025): 224–231, https://doi.org/10.70670/sra.v3i1.301.

[54] D. Hooshyar and Y. Yang, "Problems With SHAP and LIME in Interpretable AI for Education: A Comparative Study of Post-hoc Explanations and Neural-Symbolic Rule Extraction," *IEEE Access* 12 (2024): 137472–137490, https://doi.org/10.1109/access.2024.3463948.

[55] M. Mesinovic, P. Watkinson, and T. Zhu, "Explainable AI for Clinical Risk Prediction: A Survey of Concepts, Methods, and Modalities," *arXiv Preprint arXiv:230808407* (2023).

[56] M. N. Alam, M. Kaur, and M. S. Kabir, "Explainable AI in Healthcare: Enhancing Transparency and Trust upon Legal and Ethical Consideration," *International Journal of Engineering Research & Technology* 10, no. 6 (2023): 1–9.

[57] M. Shao, H. Yang, R. Du, et al., "Family Resilience in Cancer Treatment and Key Influencing Factors: A Systematic Review," *European Journal of Oncology Nursing* 66 (2023): 102403, https://doi.org/10.1016/j.ejon.2023.102403.

[58] C. E. Loprinzi, K. Prasad, D. R. Schroeder, and A. Sood, "Stress Management and Resilience Training (SMART) Program to Decrease Stress and Enhance Resilience Among Breast Cancer Survivors: A Pilot Randomized Clinical Trial," *Clinical Breast Cancer* 11, no. 6 (2011): 364–368, https://doi.org/10.1016/j.clbc.2011.06.008.

[59] L. Xiang, H. Wan, and Y. Zhu, "Effects of Cognitive Behavioral Therapy on Resilience Among Adult Cancer Patients: A Systematic Review and Meta-Analysis," *BMC Psychiatry* 25, no. 1 (2025): 204, https://doi.org/10.1186/s12888-025-06628-3.

[60] L. Northouse, A. Schafenacker, K. L. Barr, et al., "A Tailored Web-Based Psychoeducational Intervention for Cancer Patients and Their Family Caregivers," *Cancer Nursing* 37, no. 5 (2014): 321–330, https://doi.org/10.1097/ncc.0000000000000159.