Machine Learning with Applications 21 (2025) 100688

Contents lists available at ScienceDirect

Machine Learning with Applications

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

Hierarchical data modeling: A systematic comparison of statistical, tree-based, and neural network approaches

Marzieh Amiri Shahbazi *, Nasibeh Azadeh-Fard

Kate Gleason College of Engineering, Rochester Institute of Technology, 1 Lomb Memorial Drive, Rochester, NY 14623, USA

ARTICLE INFO

Keywords:
Hierarchical modeling
Multilevel analysis
Tree-based methods
Neural networks
Statistical models
Comparative analysis
Healthcare data
Nested structures
Model selection
Computational efficiency

ABSTRACT

Hierarchical modeling approaches have evolved significantly, yet comprehensive comparisons between fundamentally different methodological paradigms remain limited. This research presents a systematic comparative analysis of three distinct hierarchical modeling approaches: statistical (Hierarchical Mixed Model), tree-based (Hierarchical Random Forest), and neural (Hierarchical Neural Network). Based on the 2019 National Inpatient Sample — comprising more than seven million records from 4568 hospitals across four U.S. regions — the models were assessed for their ability to predict length of stay at the patient, hospital, and regional levels. The evaluation framework integrated quantitative metrics and qualitative factors, including analyses across varying sample sizes, simplified hierarchies, and a separate intensive-care dataset. Results demonstrate that tree-based approaches consistently outperform alternatives in predictive accuracy and explanation of variance while maintaining computational efficiency. These performance patterns remain generally consistent across sample sizes, simplified hierarchies, and the external dataset. Neural approaches excel at capturing group-level distinctions but require substantial computational resources and exhibit prediction bias. Statistical approaches offer rapid inference and interpretability but underperform in accuracy at intermediate hierarchical levels. Each model exhibits distinctive hierarchical information processing: neural models favor bottom-up flow, statistical models emphasize top-down constraints, and tree-based models achieve balanced integration. This research establishes practical guidelines for selecting appropriate hierarchical modeling approaches based on data characteristics, computational constraints, and analytical requirements, thereby advancing understanding of fundamental trade-offs in multilevel analysis.

1. Introduction

The increasing complexity of modern data analysis has revealed a key limitation in traditional models: their inability to accurately represent the complex relationships inherent in nested data structures, leading to flawed insights and decisions (Dowding & Haufe, 2018; McNabb & Murayama, 2021). Non-hierarchical (flat) models, encompassing a wide range of statistical and machine learning approaches, treat data as a single homogeneous level without considering group structure (Gelman, 2007). These models focus solely on direct relationships between predictors and outcomes, making them suitable for independent data but inadequate when observations are correlated within groups (Raudenbush, 2002; Snijders & Bosker, 2011).

Hierarchical or multilevel models address this limitation by explicitly accounting for nested structures through the incorporation of fixed effects to capture overall trends and random effects to account for group-specific variability (Bafandeh et al., 2018; Baheri, 2025; Tessler, 2014; Vermunt & Magidson, 2005). These models excel at analyzing data with natural hierarchical structures where observations are grouped or clustered, such as when multiple measurements are collected from the same individuals over time or when data comes from employees nested within different departments and organizations (Goldstein, 2011; Moen et al., 2016; Zyzanski et al., 2004). By simultaneously handling multiple levels of variation, hierarchical models provide more accurate effect estimations at different levels of the hierarchy (Dowding & Haufe, 2018; Vermunt & Magidson, 2005), resulting in more reliable estimates and reduced risk of Type I errors that typically occur when nested structures are ignored (Dowding & Haufe, 2018; McNabb & Murayama, 2021).

The adoption of hierarchical modeling has become increasingly crucial across psychology, healthcare, neuroscience, and organizational research, where nested data structures are prevalent (McNabb & Murayama, 2021; Tessler, 2014; Vermunt & Magidson, 2005). Beyond improving statistical inference accuracy, these methods enable the

* Corresponding author.
E-mail addresses: ma7684@rit.edu (M.A. Shahbazi), nafeie@rit.edu (N. Azadeh-Fard).

https://doi.org/10.1016/j.mlwa.2025.100688
Received 10 April 2025; Received in revised form 28 May 2025; Accepted 9 June 2025
Available online 24 June 2025

examination of cross-level interactions and contextual effects that traditional flat modeling approaches might miss (Dowding & Haufe, 2018; Tessler, 2014; Vermunt & Magidson, 2005). Recent interdisciplinary work further illustrates this trend, including the application of hierarchical frameworks to multi-level genomic and proteomic data in personalized oncology (Tambe-Jagtap & Jaaz, 2023), and the use of hierarchy-aware AI methods in cybersecurity (Al-Rubaye & Türkben, 2024). While these applications differ in domain, they share a structural modeling perspective that underscores the broad relevance of hierarchical approaches across fields.

Different hierarchical modeling paradigms address data hierarchy through distinct approaches. Statistical hierarchical models explicitly incorporate multiple levels of analysis to account for nested structures (Asampana Asosega et al., 2024; Dowding & Haufe, 2018). Tree-based models possess an inherent hierarchical structure in their decision-making process, naturally accommodating nested relationships (Breiman et al., 1984) while also explicitly capturing distinctions between hierarchical levels. Neural networks, though not inherently hierarchical, can be designed with deep architectures that learn hierarchical data representations (Ma et al., 2018). Each approach offers unique capabilities in handling complex hierarchical data structures.

The evolution of hierarchical modeling has led to the development of diverse methodological approaches with complementary strengths. Statistical hierarchical models excel in capturing uncertainty and providing interpretable parameter estimates (Wikle, 2016), while tree-based hierarchical methods offer flexible non-parametric solutions for complex hierarchical patterns (Salditt et al., 2023). Hierarchical neural networks have further expanded the methodological toolkit by introducing deep learning capabilities (Zhang et al., 2021). Despite these advancements, comprehensive comparisons between these methodologically distinct approaches remain limited, particularly in controlled settings using identical data structures.

Individual studies have extensively explored specific model types. Hierarchical–statistical models, such as Bayesian hierarchical approaches, have demonstrated improved forecasts and estimates across various domains (de Resende & Alves, 2020; Li et al., 2023; Zaidan et al., 2015). Hierarchical tree-based methods have evolved to incorporate sophisticated clustering techniques with applications in customer segmentation, gene expression analysis, and image processing (Corigliano et al., 2021; Hooda, 2017; Munmun & Khatun, 2022). Hierarchical neural network architectures have been implemented across fields, leveraging deep learning for complex hierarchical data representation (Guo et al., 2020; Virupakshappa et al., 2018; Yuan et al., 2020). Evaluation frameworks for these models have focused on performance metrics, comparative methodologies, benchmark datasets, and evaluation criteria (Bojic et al., 2023; Ebnehoseini et al., 2021; Zeng et al., 2021).

Prior research has extensively compared hierarchical models with non-hierarchical approaches (Chen et al., 2016; Harbord et al., 2008; Wikle, 2019), demonstrating the superior performance of hierarchical models for naturally grouped observations in education, healthcare, and social sciences (Fan et al., 2021; Heck & Thomas, 2015; Hox et al., 2017). Their effectiveness in parameter estimation, especially with imbalanced data, addresses key limitations of flat models (Thrane & Talbot, 2019; Xiao et al., 2023). However, comparative analyses between different hierarchical frameworks have primarily focused on application-specific performance. Studies comparing Bayesian models with random forests highlight trade-offs between probabilistic inference and scalability in high-dimensional datasets (Opoku Larbi et al., 2024), but comprehensive comparisons across statistical, tree-based, and neural frameworks remain limited. This gap particularly concerns relative strengths in scalability, robustness, and applicability to big data (Guo et al., 2021), creating opportunities for systematic evaluation of these architectures.

In addition to the statistical, tree-based, and neural hierarchical models evaluated in this research, numerous alternative hierarchical approaches have been proposed in the literature. These include hierarchical reinforcement learning for multi-level decision-making tasks (Zhao et al., 2023), hierarchical clustering methods for discovering nested structures in unsupervised settings (Ritzert et al., 2025; da Silva Goncalves et al., 2024), and hierarchical topic models for text analysis (Koltcov et al., 2021), among others. While these methods contribute valuable tools within their respective domains, they fall outside the scope of this study: clustering and topic models are typically unsupervised and not designed for predictive tasks, and reinforcement learning operates in sequential environments requiring task-specific reward structures. This study focuses on supervised, general-purpose hierarchical models commonly used in structured prediction problems across domains such as healthcare, education, and economics. The selection includes one representative from each of three major modeling paradigms to enable a comprehensive yet focused comparison. These models reflect diverse trade-offs in interpretability, scalability, and modeling flexibility, allowing for systematic evaluation under a unified experimental framework.

This research aims to conduct a comprehensive comparative analysis of three distinct hierarchical modeling approaches: Hierarchical–statistical, Hierarchical tree-based, and Hierarchical Neural Networks. The primary objective is to evaluate how these fundamentally different architectural paradigms handle complex nested structures and multi-level relationships within data. Through rigorous empirical investigation that considers different sample sizes among other factors, this study seeks to quantify performance differences, computational efficiency, and model interpretability, while establishing systematic criteria for model selection by examining how each approach manages trade-offs between precision, interpretability, and computational demands in hierarchical learning tasks.

Research Contributions: This study makes several significant contributions to the field of hierarchical modeling. First, it provides the first systematic comparison of statistical, tree-based, and neural network approaches to hierarchical learning, filling a critical gap in understanding how different mathematical frameworks handle nested structures. Second, it introduces an enhanced evaluation framework that integrates both quantitative metrics (accuracy, computational efficiency) and qualitative factors (interpretability, scalability), while examining performance across different sample sizes, providing a more comprehensive assessment across different hierarchical modeling paradigms. While existing evaluations are often application-specific, this framework enables a broader, more systematic comparison. Third, it establishes practical guidelines for practitioners to select the most appropriate hierarchical modeling approach based on their specific data characteristics and requirements. These findings will enable researchers and practitioners to make informed decisions about model selection, thereby advancing our understanding of hierarchical modeling capabilities across various architectural paradigms.

2. Methodology

2.1. Data description and preprocessing

This study utilizes the 2019 National Inpatient Sample (NIS) from the Healthcare Cost and Utilization Project (HCUP), the largest all-payer inpatient care database in the United States. A subset of 7,083,805 inpatient records was selected from 4568 hospitals across four regions, containing 18 key variables related to patient demographics, diagnoses, procedures, and hospital characteristics (teaching status, bed size, and urban/rural status), as well as hospital region. These variables served as input features for all hierarchical models evaluated in the study. The primary prediction target, Length of Stay (LOS), serves as a key metric for the utilization of healthcare resources (Table 1).

To improve data quality and model reliability, patients with outpatient status, neonatal status, and extreme outliers were excluded, ensuring that LOS predictions reflect typical hospital stays. Outliers were removed from the target variable (LOS) using the Interquartile

Table 1
Input and output variables across hierarchical levels.

Level	Variable	Description and statistics
Input: Patient	Age	Patient age in years (Range: 1–90, Mean: 56.6, SD: 21.9)
	Mortality Status	In-hospital mortality indicator (0 = Survived: 98.0%, 1 = Died: 2.0%)
	Admission Type	Admission status (0 = Non-elective: 77.9%, 1 = Elective: 22.1%)
	Gender	Patient gender (0 = Male: 43.0%, 1 = Female: 57.0%)
	Injury Status	ICD-10-CM injury diagnosis (0 = No injury: 90.7%, 1 = Primary injury: 5.6%, 2 = Secondary injury: 3.7%)
	Diagnosis Count	Number of ICD-10 diagnoses recorded (Range: 0–40, Mean: 13.3, SD: 7.2)
	Procedure Count	Number of ICD-10 procedures performed (Range: 0–25, Mean: 1.8, SD: 2.4)
	Service Line	Clinical service line (1 = Medicine: 12.5%, 2 = Surgery: 6.1%, 3 = Maternal: 5.2%, 4 = Psychiatric: 22.9%, 5 = Trauma: 53.2%)
	Payer	Primary expected payer (1 = Medicare: 46.5%, 2 = Medicaid: 19.4%, 3 = Private: 26.8%, 4 = Self-pay: 4.1%, 5 = No charge: 0.3%, 6 = Other: 2.9%)
	Race/Ethnicity	Patient race/ethnicity (1 = White: 67.4%, 2 = Black: 15.1%, 3 = Hispanic: 11.2%, 4 = Asian/PI: 2.7%, 5 = Native American: 0.7%, 6 = Other: 2.9%)
	Income Quartile	ZIP code median income quartile (1 = Lowest: 31.9%, 2 = Second: 25.0%, 3 = Third: 23.9%, 4 = Highest: 19.3%)
	Mortality Risk	APR-DRG risk of mortality (0 = No class: 0.0%, 1 = Minor: 47.1%, 2 = Moderate: 24.4%, 3 = Major: 19.6%, 4 = Extreme: 8.9%)
	Illness Severity	APR-DRG severity of illness (0 = No class: 0.0%, 1 = Minor: 28.5%, 2 = Moderate: 39.8%, 3 = Major: 22.6%, 4 = Extreme: 9.2%)
	Comorbidities	Presence of Elixhauser comorbidities (0 = No: 69.1%, 1 = Yes: 30.9%)
Input: Hospital	Bed Size	Hospital bed size category (1 = Small: 22.1%, 2 = Medium: 28.7%, 3 = Large: 49.1%)
	Teaching Status	Hospital location/teaching status (1 = Rural: 8.5%, 2 = Urban non-teaching: 17.7%, 3 = Urban teaching: 73.7%)
	Ownership	Hospital ownership type (1 = Government: 11.5%, 2 = Private non-profit: 74.5%, 3 = Private for-profit: 14.0%)
Input: Region	Geographic Region	Hospital census region (1 = Northeast: 18.4%, 2 = Midwest: 22.3%, 3 = South: 39.8%, 4 = West: 19.5%)
Output	Length of Stay	Duration of hospitalization in days (Range: 1–365, Mean: 4.9, SD: 6.6)

Range (IQR) method, eliminating values beyond 1.5 times the IQR. Specifically, LOS values below 1 day or above 23 days were excluded, removing 98,616 records (approximately 1.6%) from the dataset. The 10th–90th percentile range was also reviewed to assess distributional skewness caused by extreme hospital stays. Missing data were handled using simple imputation: for categorical variables, the mode was used, and for numerical variables, the median. This approach was selected due to the relatively low proportion of missingness (<2%) and the large dataset size, where complete-case analysis would have resulted in unnecessary data loss. Multiple imputation was considered, but deemed unnecessary as patterns of missingness appeared random and imputation diagnostics showed stable results across folds. Multicollinearity among predictors was assessed using Variance Inflation Factor (VIF) analysis and correlation tests.

As this study uses publicly available, de-identified data, it is exempt from institutional review board (IRB) approval and does not require informed patient consent.

2.2. Model specifications

Three hierarchical models were selected to examine different approaches to handling hierarchical data, each representing a distinct paradigm for managing multilevel structures. The Hierarchical Mixed Model (HMM) utilizes mixed-effects modeling to account for variability across levels, the Hierarchical Random Forest (HRF) captures hierarchical patterns through tree-based grouping, and the Hierarchical Neural Network (HNN) leverages deep learning embeddings to model hierarchical dependencies. In the following, we provide a detailed discussion of each model.

(a) **Hierarchical Mixed Model (HMM):** HMM addresses hierarchy by incorporating random effects to capture variations at different hierarchical levels. In this study, LOS serves as the response variable, with patients, hospitals, and regions forming different hierarchical levels. This model accounts for both fixed effects (common across groups) and random effects (specific to hierarchical levels), ensuring a structured analysis of variability across levels.

The general formulation of the mixed-effects model is:

$$LOS_{ijk} = X_{ijk}\beta + Z_{ijk}u_k + Z_{ij}u_j + Z_i u_i + \epsilon_{ijk}, \quad (1)$$

where:

- LOS_{ijk} represents the Length of Stay for patient i in hospital j within region k.
- X_{ijk} is the design matrix containing predictor variables across all hierarchical levels.
- β is the vector of fixed-effect coefficients estimating the average relationship between predictors and outcome.
- Z_{ijk}, Z_{ij}, and Z_i are the design matrices for random effects at region, hospital, and patient levels, respectively.
- u_k, u_j, and u_i are the random-effect coefficient vectors for each level, each following a normal distribution $N(0, \sigma^2)$.
- ϵijk is the residual error that captures unexplained variability that is not accounted for by fixed or random effects.

The model was implemented using Python's statsmodels library, leveraging its MixedLM class for hierarchical modeling. Estimation was performed using Restricted Maximum Likelihood (REML) with the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm as the primary optimizer. BFGS is a quasi-Newton method that approximates the Hessian matrix using gradient evaluations, offering efficient convergence for smooth, unconstrained optimization problems (Nocedal & Wright, 2006). However, in hierarchical models with complex variance structures, the likelihood surface may contain regions where the Hessian approximation becomes ill-conditioned, potentially causing convergence failures (Lindstrom & Bates, 1988). To address this challenge, a fallback strategy to the Nelder–Mead simplex algorithm was implemented. Unlike BFGS, Nelder–Mead is a derivative-free method that can navigate irregular likelihood surfaces more robustly, albeit at slower convergence rates (Lagarias et al., 1998). This dual-optimizer approach ensures reliable parameter estimation across diverse data scenarios, even in complex hierarchical datasets with large-scale healthcare data.

This model accounts for intragroup correlations and allows shrinkage estimation, improving reliability for small sample sizes within groups. Using mixed-effects modeling, the HMM captures individual-level variations and hierarchical dependencies within the hospital and regional structure, providing a comprehensive framework for analyzing LOS variations.

(b) **Hierarchical Random Forest (HRF):** This model sequentially captures the nested structure of healthcare data using multiple levels of random forests. It first predicts LOS using patient-specific characteristics. Next, it refines these predictions by modeling the residuals — unexplained variations from the patient-level model — using hospital-level features, capturing hospital effects. Finally, the residuals from the hospital-level model are further refined by incorporating regional features, which account for broader geographic trends. This stepwise approach ensures that predictions progressively incorporate effects from all levels of the hierarchy. The Fig. 1 shows the architecture of HRF. This stepwise hierarchical design, as opposed to a single-level random forest with all features, ensures that the model properly accounts for the nested data structure by sequentially modeling effects at their appropriate hierarchical levels.

The model was implemented using Python's scikit-learn library, with hyperparameters optimized via 5-fold cross-validation and grid search. The following parameter ranges were explored: number of trees 50, 75, 100, max depth 10, 12, 15, and min samples per leaf 3, 5. The final model used 100 trees (depth=15) at the patient level, 75 trees (depth=12) at the hospital level, and 50 trees (depth=10) at the regional level, all with a minimum of 5 samples per leaf for stability.

(c) **Hierarchical Neural Network (HNN):** This deep learning model captures the nested structure of healthcare data by incorporating patient, hospital, and regional effects. It processes patient-specific features through a dedicated network while integrating hospital- and region-level variables to account for hierarchical influences. Regional variables are processed first to capture broad geographic effects, which are then refined to provide hospital-level representations. These hierarchical representations, combined with patient features, serve as the final input for prediction, as shown in Fig. 2. Unlike a standard neural network where all features are processed in a single layer, this hierarchical architecture utilizes dedicated embedding networks for each level, creating specialized representations that preserve the hierarchical relationships between patients, hospitals, and regions.

The model was implemented using Python's PyTorch library, with hyperparameters tuned via 5-fold cross-validation and grid search. The grid included: learning rate 0.001, 0.0005, 0.0001, batch size 64, 128, hidden layers (64, 32), (128, 64), dropout rate 0.0, 0.1, 0.2, and weight decay 0.0, 0.001, 0.01. The selected model used a learning rate of 0.0001, batch size of 128, two hidden layers with 64 and 32 neurons, dropout of 0.1, and weight decay of 0.01. Early stopping with a patience of 5 was applied over a maximum of 30 epochs.

2.3. Training and evaluation pipeline

To rigorously compare hierarchical models, a combination of quantitative and qualitative evaluation metrics was used. These metrics were chosen to provide a comprehensive assessment of the predictive accuracy, reliability, and hierarchical structure handling capabilities of each model.

The dataset was partitioned with an 80/20 training-test split. For hyperparameter optimization, we employed 5-fold cross-validation on the training set—a statistically sound approach equivalent to using multiple validation sets (Hastie et al., 2009; Stone, 1974). This methodology maximizes available training data while preventing information leakage from the test set. All reported metrics were calculated exclusively on the held-out test set, with bootstrap resampling (50 iterations) providing confidence intervals. The partitioning maintained proportional representation across hierarchical levels, ensuring proper evaluation of multi-level effects.

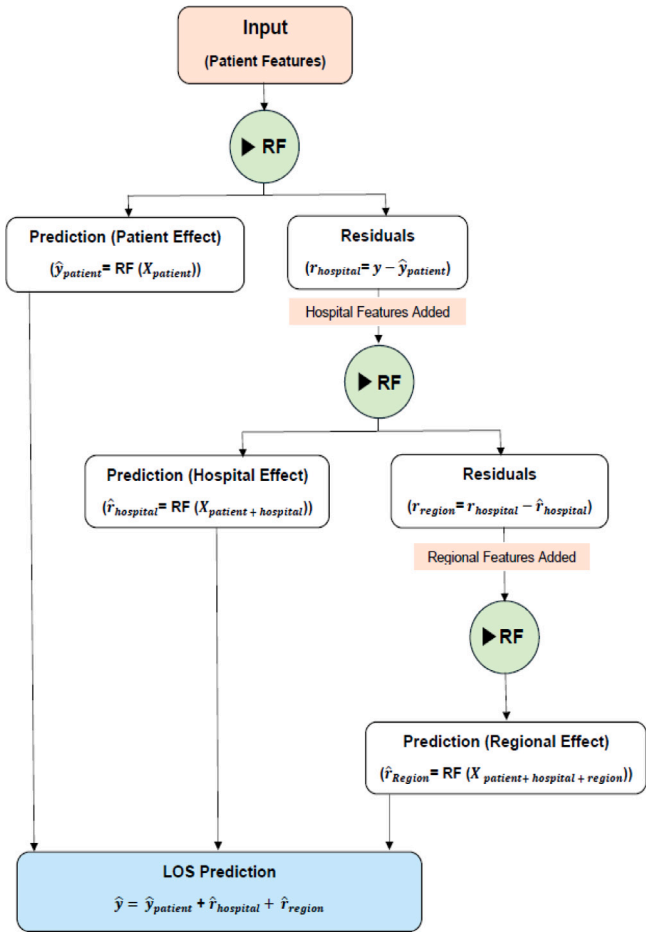


Fig. 1. HRF architecture.

Performance metrics provided a quantitative assessment of predictive accuracy and computational efficiency. This research specifically utilized the mean squared error (MSE) to quantify the average squared deviation between the observed and predicted target (LOS), providing a direct measure of predictive accuracy. The coefficient of determination (R^2) was used as a standardized metric to assess the proportion of variance explained by each model. Standard deviations (SD) were also calculated at each level (patient, hospital, and regional) to quantify uncertainty and robustness at hierarchical levels. In addition, training and testing time was measured and compared for each model to evaluate computational efficiency and scalability.

Error analysis included examining bias (mean residuals), mean absolute errors (MAE), and residual variance distributions at each hierarchical level. This approach facilitated an in-depth understanding of model accuracy and the consistency of error distribution, highlighting potential systematic biases at different hierarchical levels.

To evaluate each model's capability to manage hierarchical data with precision, several metrics were analyzed. The Intraclass Correlation Coefficient (ICC) measured the proportion of variance in the outcome that was explained by group differences at each hierarchical level—for example, between hospitals at the hospital level or between regions at the regional level. The shrinkage factor evaluated the degree of stabilization in group-level estimates by pulling them toward the overall mean. Entropy metrics provided further insight into predictive uncertainty across hierarchical levels. Overall entropy measured total prediction uncertainty, reflecting variability or dispersion in predictions, while conditional entropy quantified the remaining uncertainty at a lower hierarchical level after considering a higher

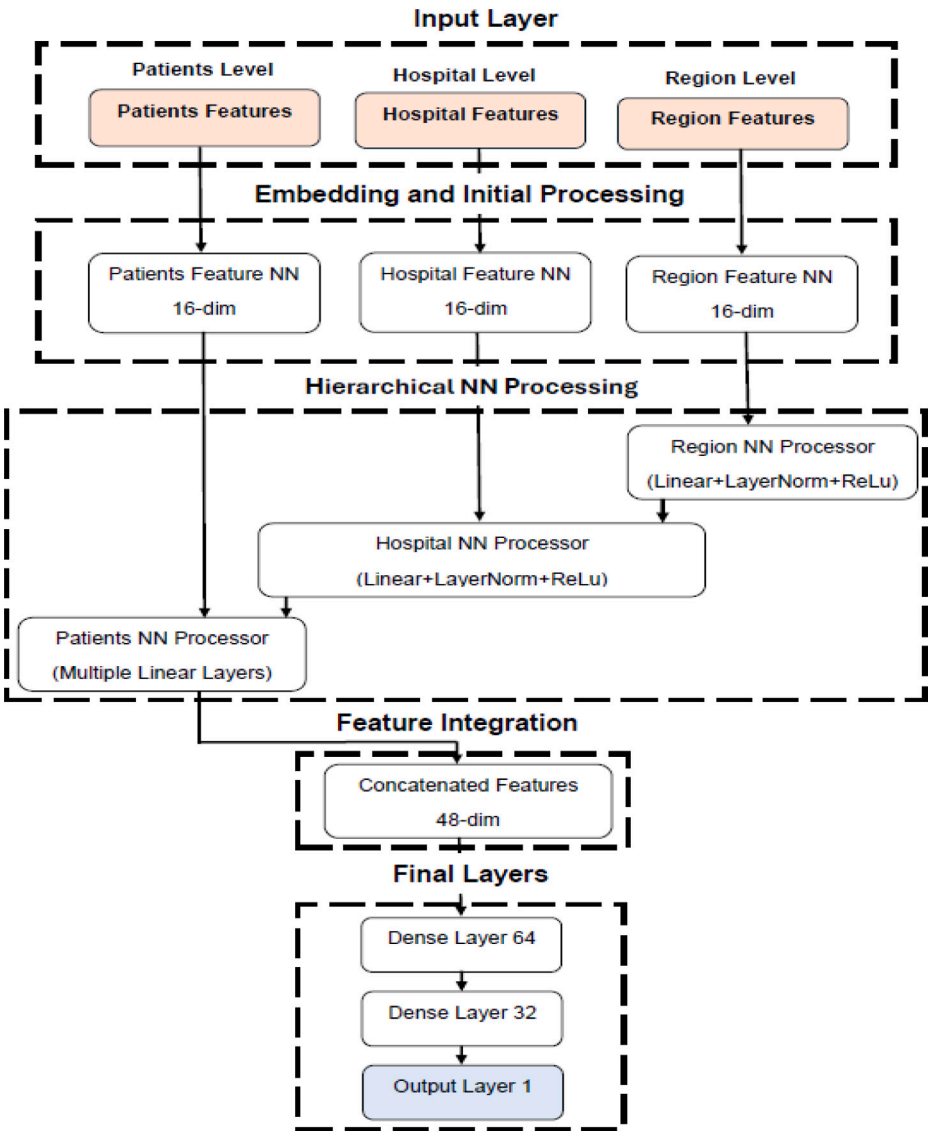


Fig. 2. HNN architecture.

level, thus clarifying how effectively higher-level group information reduced uncertainty at lower hierarchical levels.

Information flow metrics, measured both bottom-up and top-down, indicate how variance propagates through the hierarchical structure of the data (see Snijders & Bosker, 2011). Bottom-up flow quantified how patient-level information accumulates and propagates upward to form distinguishable patterns at higher levels—essentially measuring whether individual patient differences successfully translate into meaningful hospital-level and regional-level distinctions. The top-down flow demonstrated how higher-level groups, such as regions, influence or shape predictions at lower levels, like hospitals, capturing the constraining effect of broader geographical factors. Together, these metrics helped explain the direction and dynamics of variance movement and demonstrated how effectively the models capture hierarchical relationships. Table 2 provides the formulas for each metric used to address the hierarchical structure of different models.

Cross-level correlation metrics assessed relationships between hierarchical levels (patient–hospital, hospital–region, patient–region), clarifying how effectively each model integrated hierarchical information into its predictions.

Reliability reflects the consistency and reproducibility of model predictions. This research assessed reliability through model stability, measuring prediction consistency across repeated analyses or different data subsets. Additionally, variance decomposition was used to evaluate the proportion of variance explained versus unexplained by each model, highlighting how effectively each model captured hierarchical structures and data relationships.

Additionally, to investigate model sensitivity to sample size, this research analyzed model performance and reliability across three different dataset sizes: large, medium, and small. The large sample comprised the entire dataset, with approximately 7,083,805 patients, and 4568 hospitals across four regions. The medium sample consisted of 50% of the original dataset, maintaining the same number of hospitals and regions as the overall dataset. The small sample comprised approximately 5% of the data, also preserving the same number of hospitals and regions. Performance across these varying sample sizes was evaluated using MAE and R^2 distribution, while reliability was examined through variance decomposition and model stability. This sensitivity analysis provided critical insights into model robustness and scalability.

2.4. Sample size sensitivity design

To assess model robustness and scalability, this study conducted a sensitivity analysis by evaluating performance across three different

Table 2
Key hierarchical metrics and their formulas.

Metric	Formula
ICC	$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}$
Shrinkage Factor	$\frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \frac{\sigma^2_{\text{within}}}{n_g}}$ where n_g is the sample size of group g
Entropy (Overall)	$H(X) = -\sum_i P(x_i) \log P(x_i)$ where $P(x_i)$ is the probability of outcome x_i, measuring total uncertainty in variable X
Conditional Entropy	$H(X\|Y) = H(X) - I(X;Y)$ measuring remaining uncertainty in X after knowing Y
Bottom-Up Information Flow	$I_{\text{bottom-up}} = I(\text{hospital}; \text{patient})$ where $I(A;B)$ measures how much patient-level information improves hospital-level predictions
Top-Down Information Flow	$I_{\text{top-down}} = I(\text{hospital}; \text{region})$ measuring how much regional information constrains hospital-level outcomes

dataset sizes. The large sample used the full dataset, comprising approximately 7,083,805 patient records from 4,568 hospitals across four regions. The medium sample consisted of 50% of the original dataset, preserving the full hospital and region coverage. The small sample included approximately 5% of the data, again retaining representation across all hospitals and regions.

Each sample was used to train and test all three models (HMM, HRF, and HNN) using the same pipeline described earlier. Performance was evaluated using R^2 and MAE, while model reliability was examined using variance decomposition and model stability scores. This design allowed us to isolate the effects of data volume on predictive accuracy, hierarchical signal propagation, and consistency of model behavior across scales.

2.5. Hierarchy sensitivity analysis design

To examine how hierarchical model performance is influenced by the structure and depth of the hierarchy, a controlled sensitivity analysis conducted by systematically altering the levels of nesting within the dataset. Specifically, two reduced-hierarchy configurations evaluated:

1. **Region-Only Hierarchy:** The hospital level was removed, retaining only the patient-to-region structure. This simplification tests the ability of models to capture top-down variation in the absence of intermediate institutional grouping.
2. **Hospital-Only Hierarchy:** The region level was excluded, preserving the patient-to-hospital relationship. This configuration emphasizes bottom-up information flow, modeling institutional variation without higher-level regional aggregation.

In both configurations, all models (HMM, HRF, and HNN) were retrained using the same preprocessing pipeline, hyperparameters setting, and evaluation metrics as in the full three-level hierarchy. Group identifiers for the retained levels were re-encoded into normalized integer indices, and patient-level features remained unchanged. This experimental design enabled us to isolate the contribution of each hierarchical level and assess the robustness of each model to structural simplification.

2.6. External validation dataset

To evaluate the generalizability of the findings, external validation conducted using a structured clinical dataset derived from the publicly available MIMIC-IV critical-care database (Johnson et al., 2023). The cohort consisted of 100 patients, nested within 9 ICU units across 11 hospital departments, forming a 3-level hierarchical structure (patients → ICU units → departments). The prediction target was ICU length of stay, a continuous measure of resource utilization during critical care episodes. The dataset included demographic, clinical, and administrative variables, namely: age, gender, race, insurance type, admission type, number of diagnoses, number of procedures, presence of comorbidities, and in-hospital mortality. All features were encoded appropriately: categorical variables were label-encoded, and hierarchical group identifiers were normalized into unique integer indices. No missing values were present. To ensure consistency with the primary dataset (NIS 2019), we applied the same modeling pipeline, including data preprocessing, train-test splitting (80/20), and evaluation metrics (R^2 and MAE). This dataset allowed us to assess model robustness in a distinctly different clinical context, using ICU-level operational granularity instead of hospital-level aggregates.

3. Results

To investigate how different hierarchical modeling approaches capture the structure within the data, the performance of three fundamentally distinct models was analyzed. The comparative analysis prioritizes the models' hierarchical capabilities ensuring that the insights directly reflect methodological distinctions.

The analysis examined three hierarchical levels: lower (individual/patient), middle (group/hospital), and upper (systemic/region). The Results section is structured into the following distinct parts: (1) Performance Analysis, (2) Error Analysis, (3) Hierarchical Structure Analysis, (4) Multi-level Effects, (5) Reliability Analysis, and (6) Sensitivity to Sample Size.

3.1. Model performance comparison

Comprehensive analysis reveals distinct performance patterns across the three hierarchical models when evaluated using multiple metrics. Table 3 presents the complete performance metrics for HRF, HNN, and HMM across different hierarchical levels, while Table 4 summarizes the computational efficiency of each approach. HRF consistently demonstrates superior predictive performance, achieving the highest overall R^2 value of 0.436 ± 0.001 and the lowest MSE of 7.878 ± 0.320. This performance remains remarkably consistent at both the regional level (R^2 = 0.467 ± 0.001, MSE = 7.857 ± 0.014) and patient level (R^2 = 0.466 ± 0.000, MSE = 7.888 ± 0.000). However, like all models in this study, HRF struggles at the hospital level, exhibiting a slightly negative R^2 with high variability (−0.018 ± 2.784).

HNN delivers intermediate performance with an overall R^2 of 0.391 ± 0.012 and MSE of 8.564 ± 0.002. Similar to HRF, it maintains consistent performance at the regional level (R^2 = 0.387 ± 0.000, MSE = 8.551 ± 0.000) and patient level, but experiences more pronounced deterioration at the hospital level (R^2 = −0.062 ± 2.641, MSE = 7.886 ± 7.197). The standard deviation of R^2 at the hospital level (2.641) indicates substantial prediction instability at this hierarchical

Table 3
Model performance metrics (mean ± standard deviation) across hierarchical levels.

Model	Level	R^2 (mean ± std)	MSE (mean ± std)
HRF	Overall	0.436 ± 0.001	7.878 ± 0.320
	Region	0.467 ± 0.001	7.857 ± 0.014
	Hospital	−0.018 ± 2.784	7.580 ± 7.929
	Patient	0.466 ± 0.000	7.888 ± 0.000
HNN	Overall	0.391 ± 0.012	8.564 ± 0.002
	Region	0.387 ± 0.000	8.551 ± 0.000
	Hospital	−0.062 ± 2.641	7.886 ± 7.197
	Patient	0.389 ± 0.006	8.554 ± 0.485
HMM	Overall	0.273 ± 0.016	10.316 ± 0.537
	Region	0.260 ± 0.001	10.319 ± 0.014
	Hospital	−0.250 ± 2.508	10.717 ± 13.297
	Patient	0.263 ± 0.006	10.306 ± 0.597

Table 4
Average training and prediction times (in seconds) for each model, reported as mean ± standard deviation.

Model	Training time (s)	Prediction time (s)
HRF	830.169 ± 250.286	2.207 ± 0.220
HNN	27,927.967 ± 0.001	21.125 ± 0.001
HMM	778.958 ± 33.322	0.530 ± 0.019

level. HMM exhibits the weakest overall performance (R^2 = 0.273 ± 0.016, MSE = 10.316 ± 0.537) across all metrics. While maintaining reasonable consistency between regional and patient levels, it shows significantly poorer performance at the hospital level (R^2 = −0.250 ± 2.508, MSE = 10.717 ± 13.297), with the highest variability in MSE among all models.

Computational efficiency varies dramatically across models (Table 4). HNN requires by far the most significant training resources at 27,927.967 s (7.8 h), approximately 34 times longer than HRF (830.169 ± 250.286 s) and 36 times longer than HMM (778.958 ± 33.322 s). Prediction time follows a similar pattern, with HNN (21.125 s) being significantly slower than HRF (2.207 ± 0.220 s), while HMM offers the fastest inference (0.530 ± 0.019 s).

The negative R^2 values observed at the hospital level indicate that model predictions perform worse than a mean-based baseline, suggesting difficulty in capturing mid-level variance. This consistent finding suggests a fundamental challenge in modeling hospital-level effects within the hierarchical structure, potentially due to greater variability or fewer observations per hospital compared to patient or regional levels.

3.2. Error and bias analysis

Error analysis across hierarchical levels reveals fundamental methodological differences in how each model handles multilevel data structures. At all hierarchical levels, HRF demonstrates consistently lower MAE values (averaging 1.87) and near-zero bias, indicating more accurate and stable predictions across hierarchical levels, which indicates its methodological effectiveness in maintaining prediction consistency across the hierarchical spectrum. This suggests that the sequential modeling approach with level-specific random forests successfully captures variance at appropriate levels without systematic distortion.

HNN exhibits a distinctive methodological pattern, characterized by moderate MAE values (approximately 2.05) and a persistent negative bias (−0.25 to −0.35) across all hierarchical levels. This consistent underestimation suggests a structural characteristic in how neural embeddings process hierarchical information—possibly stemming from the way embedding vectors propagate and combine information across levels. The uniformity of this bias across levels suggests that the neural architecture systematically applies a similar transformation to predictions regardless of hierarchical position.

HMM's error profile varies more substantially across hierarchical levels, exhibiting higher overall MAE values (approximately 2.21) and inconsistent bias patterns—neutral at the patient level, positive (0.31) at the hospital level, and neutral again at the regional level. This variability suggests that HMM may struggle to allocate variance consistently across the hierarchy. These inconsistencies are especially evident in performance metrics at the intermediate (hospital) level, where prediction stability is lowest.

Fig. 3 graphically illustrates these methodological distinctions, showing how error profiles change across levels. Particularly noteworthy is how variance (standard deviation) of prediction errors decreases for all models as we move from lower to higher hierarchical levels, but at different rates. HRF maintains the most consistent error reduction across levels (from 2.09 at patient level to 0.07 at region level), indicating its superior capability in appropriately assigning variance to the correct hierarchical level. This methodological advantage suggests that the sequential residual processing approach effectively captures level-specific information without contamination between levels.

These patterns reveal important methodological insights: HRF's approach to sequentially modeling residuals appears to provide the most balanced hierarchical error distribution; HNN's embedding approach creates consistent biases that persist across levels; and HMM's mixed-effects framework struggles to model level-specific effects appropriately. These results highlight the role of hierarchical structure in shaping prediction quality and reinforce the need for model selection aligned with multilevel data characteristics.

3.3. Model interpretability

To clarify the performance differences observed above, this study next examines model interpretability through SHAP analysis. SHAP (SHapley Additive exPlanations) decomposes each prediction into additive feature contributions, ensuring consistency with the model's output and allowing direct comparison of feature influence. A SHAP beeswarm plot generated from a stratified 30,000-record subsample (Fig. 4) shows one dot per patient: the dot's horizontal position is its signed SHAP value (positive values lengthen the predicted stay, negative values shorten it), while color encodes the original feature value (red = high, blue = low).

The plot reveals that patient-level acuity variables — Severity of Illness, Number of Procedures, Number of Diagnoses, and Service Line — dominate the length-of-stay prediction. The top hospital-level variable, Bed Size, appears only seventh, indicating that institutional context adds a modest adjustment after patient factors are accounted for. Socio-economic covariates (e.g., race, income quartile, primary payer) cluster near the origin, confirming their limited marginal impact in the presence of more salient clinical features.

Hence, HRF not only achieves the best predictive accuracy but also yields transparent, clinically plausible attributions: high acuity and procedure counts push stays longer (red dots on the right), whereas low acuity pulls them shorter (blue dots on the left). This interpretability advantage complements HRF's quantitative superiority and demonstrates its balanced use of patient- and hospital-level information (see Fig. 4).

3.4. Hierarchical structure analysis

Hierarchical structure analysis explores how effectively each model captures and utilizes structural relationships across different hierarchical levels. The key metrics assessed are ICC, shrinkage factor, overall and conditional entropy, bottom-up flow, and top-down flow. Table 5 summarizes these metrics across HRF, HNN, and HMM.

At the upper level, all models demonstrate minimal influence (ICC ≈ 0.005) from the highest hierarchical tier, indicating that upper-level factors contribute little to outcome variability. The high shrinkage factors. The 0.994 estimate reflects that upper-level estimates are pulled

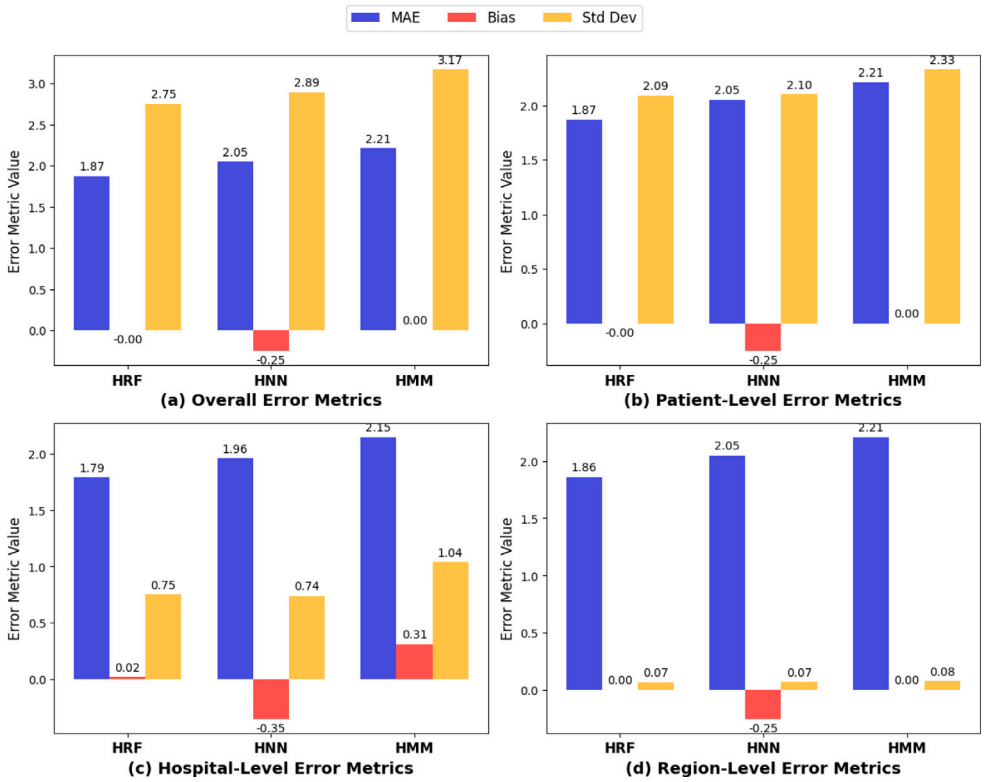


Fig. 3. Comparison of Error Metrics (MAE, Bias, Std Dev) Across Hierarchical Levels.

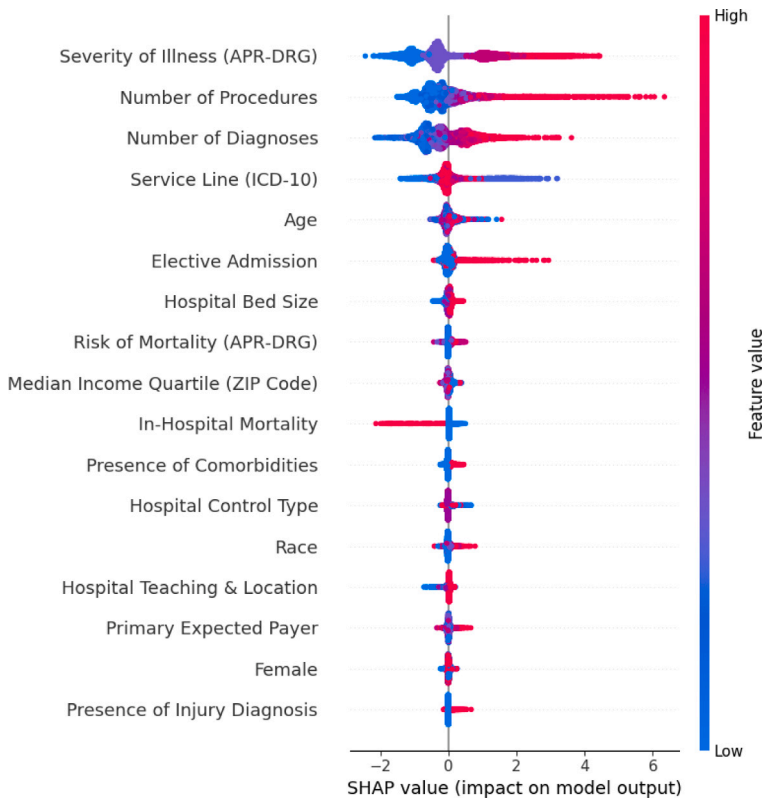


Fig. 4. SHAP beeswarm plot for the Hierarchical Random Forest (30,000-record subsample). Dots are individual patients: color shows the feature value (red = high, blue = low) and horizontal position shows that feature's positive or negative contribution to the predicted length of stay. Features are ranked top-to-bottom by total impact.

Table 5
Comparison of hierarchical metrics for HRF, HNN, and HMM.

Metric	HRF	HNN	HMM
Upper Level (Regions)			
ICC	0.005	0.006	0.005
Shrinkage Factor	0.995	0.994	0.995
Entropy	2.471	2.058	1.896
Conditional Entropy	1.001	0.998	1.001
Bottom-Up Flow	0.005	0.005	0.005
Top-Down Flow	N/A	N/A	N/A
Middle Level (Hospitals)			
ICC	0.188	0.323	0.229
Shrinkage Factor	0.812	0.677	0.771
Entropy	1.892	1.517	1.466
Conditional Entropy (Hospitals | Regions)	0.766	0.736	0.774
Bottom-Up Flow (Hospitals → Regions)	0.148	0.297	0.190
Top-Down Flow (Regions → Hospitals)	0.246	0.201	0.218
Lower Level (Patients)			
ICC	N/A	N/A	N/A
Shrinkage Factor	N/A	N/A	N/A
Entropy	0.000	0.000	0.000
Conditional Entropy (Individuals | Hospitals)	0.000	0.000	0.000
Bottom-Up Flow (Patients → Hospitals)	1.000	1.000	1.000
Top-Down Flow (Hospitals → Patients)	0.242	0.306	0.297

strongly toward the global average. Given the relationship between these metrics in multilevel modeling, the shrinkage factor approximates 1−ICC for large, balanced datasets such as ours. Despite similar ICCs, HRF captures higher upper-level entropy (2.471), suggesting it maintains more distinctiveness at this tier than HNN (2.058) or HMM (1.896), even though these distinctions contribute little to variance. The uniform bottom-up flow (0.005) confirms that minimal information travels upward from middle levels. Top-down flow is not applicable at this level because there is no higher grouping beyond the regional level.

At the middle level, the models diverge significantly in their approach to intermediate grouping structures. HNN demonstrates the strongest middle-level grouping effect (ICC = 0.323), which is substantially higher than HMM (0.229) and HRF (0.188). This is reflected in HNN's lower shrinkage factor (0.677), indicating greater preservation of group-specific deviations from the mean. Interestingly, HRF maintains higher middle-level entropy (1.892) despite its lower ICC, suggesting it preserves group distinctiveness in ways not directly reflected in variance components. HNN shows the strongest bottom-up flow from middle to upper levels (0.297), indicating it most effectively leverages intermediate patterns to inform higher-level estimates. Conversely, HRF exhibits the strongest top-down influence (0.246), showing greater upper-level constraints on middle-level predictions.

At the lower level, ICC and shrinkage metrics are marked as "N/A" because there is no deeper subdivision beneath individual units – each unit is uniquely identified in the dataset, making these metrics inapplicable. The zero-entropy values across all models confirm that individual identifiers completely determine group membership, with no remaining uncertainty. The conditional entropy being zero further confirms that knowing an individual's group entirely classifies that individual uniquely. The uniform bottom-up flow of 1.000 indicates complete information transfer from individual units to their respective groups. The most revealing metric is top-down flow, where HNN (0.306) demonstrates significantly stronger middle-to-lower information flow than HRF (0.242), with HMM (0.297) falling between them. This variances, which suggests that HNN relies more heavily on group-level characteristics when making individual predictions, whereas HRF preserves more unit-specific variances, independent of group membership.

These metrics reveal fundamental differences in hierarchical information processing across modeling approaches: HNN excels at capturing and utilizing middle-level effects, HRF maintains better distinction between levels while allowing for more independent lower-level variation, and HMM balances between these approaches with moderate information flow in both directions. The patterns of information flow and variance partitioning provide critical insights into how each model's architecture handles multi-level data structures.

Table 6
Cross-Level Correlations for HRF, HNN, and HMM.

Model	Patient–Hospital	Hospital–Region	Patient–Region
HRF	0.242	0.246	0.060
HNN	0.306	0.201	0.062
HMM	0.297	0.218	0.065

Table 7
Comparison of model reliability: stability and variance explained.

Model	Stability (Mean)	Variance explained	Variance unexplained
HRF	0.680	43.8%	56.2%
HNN	0.633	31.0%	69.0%
HMM	0.509	25.7%	74.3%

3.5. Multi-level effects

Cross-level correlations from cross-validation analyses were used to assess differences among the HRF, HNN, and HMM models regarding the strength of hierarchical associations across patients, hospitals, and regions (Table 6).

HNN exhibited the highest patient-to-hospital correlation (0.306), suggesting it effectively integrates hospital-level variation into patient-level predictions. Conversely, HRF showed the strongest hospital-to-region correlation (0.246), indicating a greater alignment between hospital-level predictions and regional-level factors compared to the other models. HMM presented balanced correlations, with a patient-to-hospital correlation (0.297) comparable to HRF (0.242) and an intermediate hospital-to-region correlation (0.218).

Patient-to-region correlations were similar across all three models (approximately 0.06), a limited direct influence from regional-level factors on individual patients in any model.

These results highlight that HNN is most sensitive to hospital-level factors when predicting individual outcomes, HRF more strongly captures regional influences on hospitals, and HMM occupies a balanced intermediate position, reflecting both individual–hospital and hospital–region linkages to a moderate extent. Overall, the cross-validation analysis clarifies that each model emphasizes hierarchical relationships differently, with HNN focusing on lower-to-middle levels, HRF emphasizing middle-to-upper levels, and HMM occupying a balanced position between these extremes.

3.6. Reliability

Reliability in hierarchical modeling refers to the stability and consistency of model predictions across different data subsets or repeated runs, which is essential for ensuring generalizability. This study evaluates the reliability of the three hierarchical models using two key metrics: model stability and variance decomposition.

Model stability measures the consistency of predictions across 50 random sampling iterations. As shown in Table 7, HRF achieved the highest stability (mean = 0.680), followed by HNN (0.633), while HMM showed the lowest stability (0.509). Although all models demonstrated relatively low variation across runs, HRF consistently outperformed the others in prediction reliability.

Variance decomposition complements the stability analysis by quantifying how much of the total outcome variance is explained by each model. HRF explains 43.8% of the variance, outperforming HNN (31.0%) and HMM (25.7%), indicating its stronger ability to capture hierarchical structure and reduce residual error.

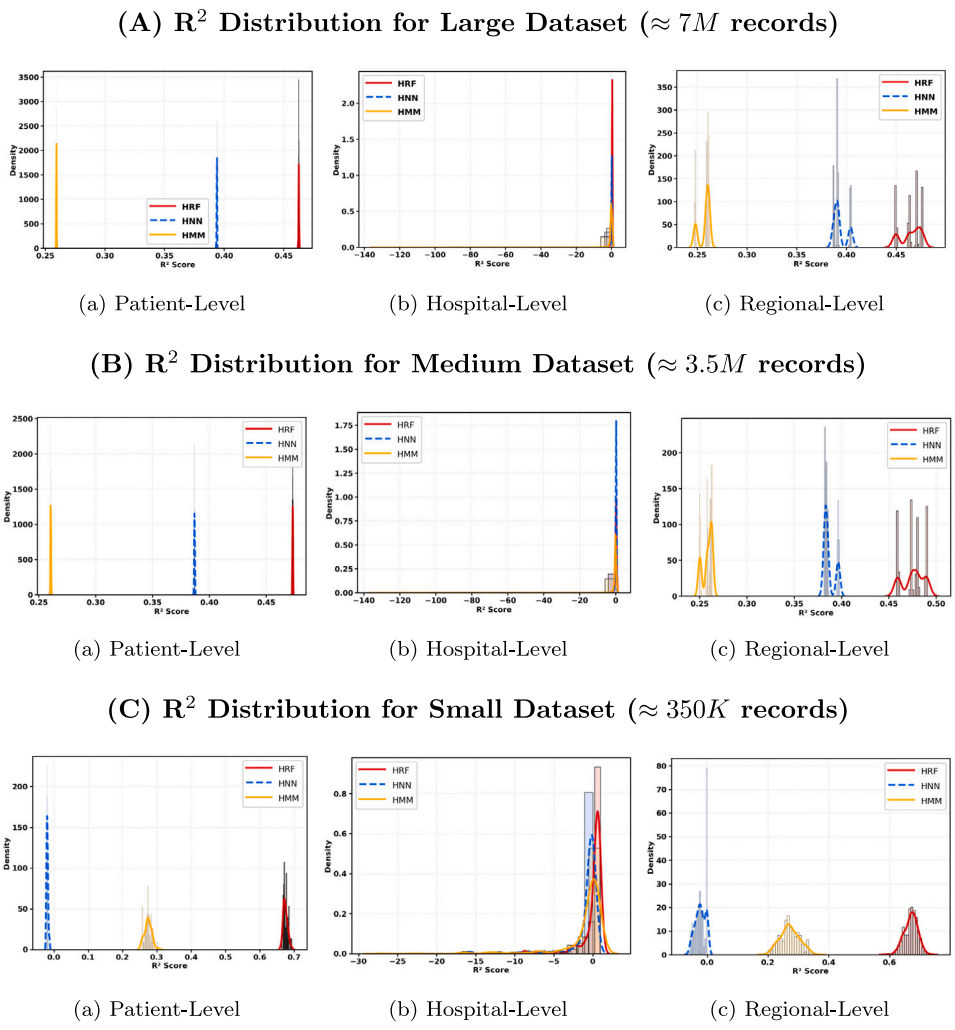


Fig. 5. Density distribution of R^2 scores across hierarchical levels (patient, hospital, and region) for different models (HRF, HNN, HMM), evaluated over varying dataset sizes. The y-axis represents the density of R^2 values, with higher peaks indicating more consistent model performance and wider distributions reflecting greater variability in predictive effectiveness.

Considering both results, the stability and variance decomposition results clearly identify HRF as the most reliable modeling approach. HNN offers moderate reliability but captures substantially less variance, while HMM, despite its balance across levels, demonstrates the weakest overall reliability. When reliability — defined by consistent predictions and effective variance partitioning — is the priority, HRF emerges as the most robust choice.

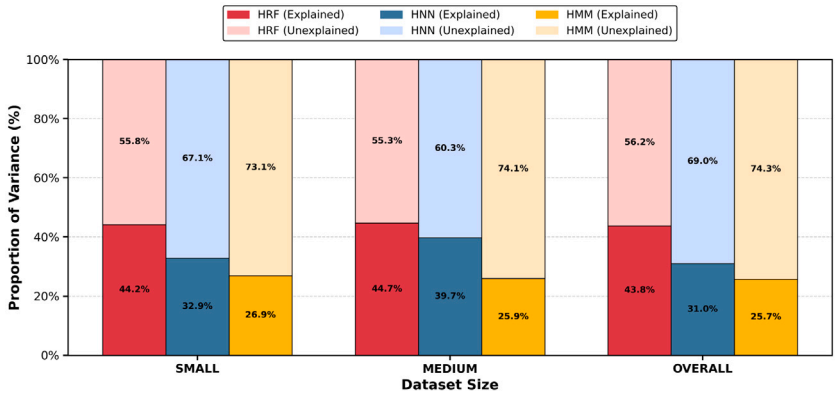
3.7. Sensitivity to sample size

Fig. 5 illustrates the density distribution of coefficient of determination (R^2) values across three dataset sizes and hierarchical levels for the HRF, HNN, and HMM models. In the large dataset (≈7M records), HRF demonstrates superior predictive performance at the patient level, with its distribution sharply peaked around $R^2 \approx 0.45$. HNN's distribution centers near $R^2 \approx 0.40$, while HMM performs substantially worse with a concentration near $R^2 \approx 0.25$. At the hospital level, all models exhibit negative R^2 values, though HNN shows the least negative distribution. At the regional level, HRF remain multiple peaks at higher R^2 values (≈0.35–0.45), HNN exhibits multiple intermediate peaks, and HMM again shows lower performance with peaks around $R^2 \approx 0.25$–0.30.

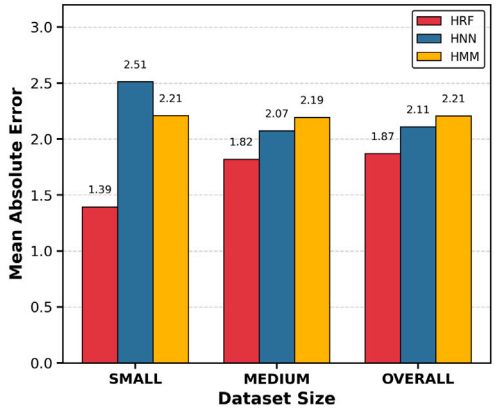
The medium dataset (≈3.5M records) follows similar trends. HRF continues to lead at the patient level ($R^2 \approx 0.45$), followed by HNN ($R^2 \approx 0.40$), while HMM remains substantially lower ($R^2 \approx 0.25$). At the hospital level, all models maintain negative R^2 values, particularly for HRF and HMM. At the regional level, the performance hierarchy is consistent with that observed in the large dataset.

The small dataset (≈350K records) reveals noteworthy shifts in distribution. At the patient level, HRF's distribution shifts rightward and broadens, peaking at $R^2 \approx 0.6$–0.7, indicating a substantial performance improvement. HNN's distribution becomes more dispersed and shifts toward lower values, while HMM shows improved relative performance with a rightward shift compared to larger datasets. At the hospital level, all models cluster near zero, with notably less negative R^2 values than in the larger datasets. Regional-level distributions in the small dataset exhibit more significant overlap between models, with HRF and HMM centering around $R^2 \approx 0.3$–0.7, and HNN displaying a bimodal pattern with peaks at lower values.

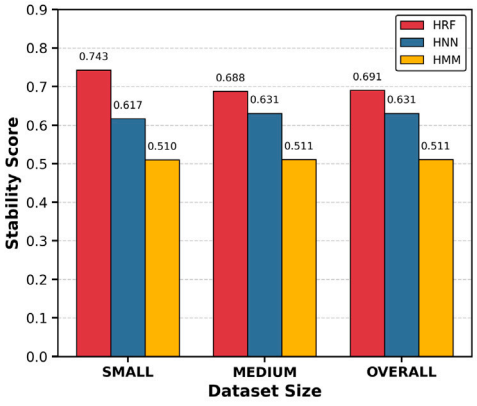
Fig. 6 provides complementary performance metrics from 3-fold cross-validation across sample sizes. HRF consistently explains the highest variance (approximately 44%) regardless of sample size, while demonstrating superior stability in small samples (0.743) and the lowest MAE (1.39–1.87). HNN exhibits notable data dependency, with explained variance improving significantly from small to medium datasets (32.9% to 39.7%) and corresponding error reduction (2.51 to 2.07). HMM maintains relatively consistent but inferior performance across metrics and sample sizes, with an explained variance of around 26%, stability measures near 0.51, and moderate error rates of around 2.2. These cross-validation findings align with the R^2 distribution patterns, confirming that HRF offers the most robust performance across dataset



((a)) Variance Decomposition



((b)) Mean Absolute Error (MAE)



((c)) Model Stability

Fig. 6. Comparison of model performance metrics across sample sizes.

Table 8
Overall performance after removing one hierarchical tier.

Model	Patient → Region		Patient → Hospital	
	R^2	MAE	R^2	MAE
HRF	0.429	1.886	0.431	1.883
HNN	0.305	2.089	0.362	2.180
HMM	0.255	2.223	0.256	2.226

Table 9
Overall performance on MIMIC-IV dataset.

Model	R^2 (mean ±std)	MAE (mean ±std)
HRF	0.216 ± 0.157	2.53 ± 0.56
HNN	0.025 ± 0.035	2.71 ± 0.71
HMM	0.195 ± 0.184	2.57 ± 0.63

sizes. At the same time, HNN benefits substantially from larger training sets, and HMM performs more competitively relative to other models when applied to smaller, potentially more homogeneous datasets.

3.8. Sensitivity to hierarchical structure

Table 8 summarizes model accuracy after retraining each algorithm under the two reduced-nesting schemes described in Section 2.5. Across both manipulations, predictive changes are marginal: for HRF the absolute difference never exceeds 0.002 in R^2 or 0.004 days in MAE, and the baselines (HNN, HMM) show similarly small shifts (all $|\Delta R^2| < 0.06$, $|\Delta \text{MAE}| < 0.10$). HRF therefore remains the top-performing model regardless of whether the intermediate *hospital* tier or the upper *region* tier is removed.

3.9. External validation

The hierarchical models were evaluated on an independent MIMIC-IV ICU dataset to assess cross-domain generalization. As shown in Table 9, HRF maintained its performance advantage, achieving an MAE of 2.53 days compared with 2.71 days for HNN and 2.57 days for HMM. The corresponding R^2 values (0.216 for HRF, 0.025 for HNN, and 0.195 for HMM) are lower than those in the primary analysis (0.43, 0.39, and 0.27, respectively)—a predictable drop given the greater clinical complexity and smaller sample size of ICU data.

Notably, while all models experienced performance degradation in this transfer scenario, HRF's relative advantage actually widened: its R^2 is roughly eight times higher than that of HNN and modestly higher than that of HMM, whereas the primary analysis showed only a two- to three-fold gap. This enhanced relative performance in a challenging, small-sample domain confirms the robustness of tree-based hierarchical approaches to model nested healthcare data structures, particularly when faced with the increased clinical complexity and volatility of the outcomes characteristic of critical care environments.

4. Discussion

The comprehensive evaluation of three distinct hierarchical modeling approaches — HRF, HNN, and HMM — reveals fundamental differences in how each architecture captures and utilizes multi-level healthcare data structures. These differences provide practical guidance for selecting models in hierarchical clinical settings.

4.1. Model architecture and predictive performance

HRF emerged as the top-performing model for hierarchical healthcare data, achieving the highest predictive accuracy ($R^2 = 0.436$), lowest error (MSE = 7.878), and superior variance explanation (43.8%). Its success can be attributed to several theoretical advantages in handling hierarchical data structures. HRF's sequential residual modeling approach effectively captures variance at each level before proceeding to the next, allowing it to identify patterns that might be obscured in simultaneous multi-level optimization. This staged approach mirrors the natural nested structure of healthcare data, where patient outcomes are influenced by progressively broader contexts (hospital, region).

HNN provided intermediate accuracy ($R^2 = 0.391$) but required extensive computational resources (27,928 s of training). Its strength lies in learning representations that strongly structure hospital-level data (ICC = 0.323) through significant bottom-up information flow (0.297). The neural embedding approach effectively distinguishes hospitals based on patient characteristics but introduces a consistent negative prediction bias across levels.

HMM, while computationally efficient (0.530 s for inference), exhibited the lowest accuracy ($R^2 = 0.273$) and explained variance (25.7%). Its predominantly top-down information flow (0.297) emphasizes regional constraints but struggles to reliably partition variance across levels. The parametric assumptions inherent in mixed models may inadequately capture the complex, non-linear relationships present in length-of-stay predictions.

4.2. The hospital-level modeling challenge

A consistent finding across all models was negative R^2 values at the hospital level, indicating poorer performance than a simple mean-based prediction at this intermediate hierarchical tier. This phenomenon suggests fundamental challenges in modeling hospital-level effects. Several factors may contribute:

- **Variable insufficiency:** While teaching status, bed size, and urban/rural status represent important hospital characteristics, they may inadequately capture the complex institutional factors influencing LOS. Critical elements such as staffing models, discharge protocols, and bed management policies remain unobserved.
- **Variance heterogeneity:** The 4568 hospitals in our dataset exhibit wide variation in patient populations and practice patterns, creating heterogeneous groupings that resist modeling with our limited set of institutional variables.
- **Cross-level interactions:** As shown in the Multi-level Effects analysis, hospital effects demonstrate substantial correlation with patient characteristics (patient–hospital correlations of 0.242–0.306 across models). This strong interdependence makes it difficult to isolate pure hospital-level effects, as their influence is partially expressed through patient-level variables rather than as independent institutional factors.
- **Overshadowing effects:** With millions of patient records containing rich individual-level predictors, and broader regional patterns established across hospital clusters, the unique contribution of hospital-specific factors may be diminished in the overall predictive framework.

Future work could employ hierarchical cross-validation methods designed for nested data — such as leave-one-hospital-out validation — to better assess how these models generalize to unseen institutions with similar characteristics. This approach could clarify whether additional hospital-level variables beyond the structural factors we examined are needed to effectively capture institutional influences on patient outcomes.

4.3. Structural considerations for hierarchical information flow

The models exhibited distinct strategies for processing hierarchical information. HRF maintained a balanced flow across levels, with an intermediate ICC (0.188), effectively preserving both group-level distinctions and individual variance. In contrast, HNN structured information in a bottom-up fashion: regional features were first processed into embeddings, passed to the hospital-level network, and then integrated with patient-level features before final prediction.

The very low regional ICC (≈ 0.005) across models indicates that regional factors contributed minimally to prediction variance. This highlights a key insight—individual and hospital-level characteristics are far more influential in determining length of stay than broader regional patterns.

4.4. Comparative strengths and limitations

Based on the results, tree-based models such as HRF offer the best general-purpose performance and scalability. HNN is ideal when understanding group-level patterns is essential, such as when hospital-level interventions are under study. HMM remains useful for quick deployment and statistical inference when interpretability is prioritized over predictive power. Although HRF can be explained post hoc through SHAP values, HMM offers intrinsic, coefficient-level interpretability with formal statistical inference, a key advantage when effect sizes and confidence intervals are required. Table 10 summarizes the key strengths and limitations of each hierarchical modeling approach as revealed by our comprehensive evaluation:

This comparative analysis reveals that specific analytical requirements, data characteristics, and computational constraints should guide the choice of hierarchical modeling approach.

4.5. Practical implications and future directions

The comparative reliability analysis highlights HRF as the most stable model, achieving the highest stability score and demonstrating robust performance even with smaller datasets, simplified hierarchies, and a separate intensive care dataset. This indicates its suitability for community hospitals or resource-constrained settings. However, HRF's batch processing approach presents challenges for online updates or incremental learning as new data becomes available, requiring periodic retraining to incorporate evolving patterns.

HNN's performance strongly depends on dataset size, improving significantly with larger samples—a typical neural network characteristic. Its computational demands, however, limit its practical deployment in time-sensitive environments.

The divergent performance patterns observed across varying sample sizes underscore the importance of considering data volume when selecting hierarchical modeling approaches. Future research should investigate whether the challenges in modeling intermediate hierarchical levels are specific to healthcare data or represent a broader phenomenon in hierarchical modeling across domains. Additionally, exploring hybrid approaches that combine the strengths of different hierarchical modeling paradigms — such as using HMM for initial variance decomposition followed by HRF for prediction — may yield further improvements.

To address these limitations, hybrid models offer a promising direction. For example, using random effects estimated from an HMM as structured inputs to HRF or gradient-boosted trees (e.g., XGBoost) could combine statistical transparency with high predictive performance. Such hybrid designs could leverage the variance decomposition capabilities of mixed models with the nonlinear modeling power of tree ensembles, potentially improving both generalization and interpretability. Alternative strategies include hierarchical stacking—combining predictions from HMM, HRF, and HNN via a meta-learner, or hierarchical XGBoost variants that sequentially refine predictions across levels of

Table 10
Comparison of hierarchical models: strengths, limitations, and suitable use cases.

Model	Key strengths	Key limitations	Best use cases
HRF	• Superior predictive accuracy • Balanced hierarchical information flow • Robust performance with limited samples • Unbiased predictions across hierarchy • Efficient computational profile • Post-hoc feature importance	• Struggles with intermediate hierarchical levels • Moderate group differentiation capacity	• General-purpose hierarchical modeling • Resource-constrained environments • Applications requiring balanced variance partitioning
HNN	• Strong capability for group-level distinction • Excellent upward information propagation • Effective feature extraction at multiple levels	• Computationally intensive • Systematic prediction bias • Requires large datasets • Slow inference process	• Data-rich environments • Applications prioritizing group distinctiveness • Cases requiring strong bottom-up pattern recognition
HMM	• Fast inference speed • Quick model training • Statistically interpretable • Balanced cross-level relationships	• Limited predictive power • Lower variance explanation capacity • Inconsistent bias distribution • Reduced stability across iterations	• Time-sensitive applications • Preliminary exploratory analysis • Research requiring statistical inference • Hypothesis testing of hierarchical effects

abstraction. These ideas merit empirical exploration to assess their scalability and fairness properties in clinical data contexts.

In addition, future work should explore bias mitigation strategies for hierarchical neural models — such as loss reweighting or adversarial debiasing — to address the systematic underestimation observed in HNN. These methods may improve fairness in clinical prediction tasks where demographic or institutional disparities are present.

As models become increasingly sophisticated, maintaining interpretability at multiple hierarchical levels will be crucial for clinical adoption. Developing visualization and explanation techniques that communicate how predictions are influenced by individual, hospital, and regional factors would significantly enhance the practical utility of hierarchical models in healthcare decision support.

5. Conclusion

This study compared statistical (HMM), tree-based (HRF), and neural (HNN) approaches within a unified evaluation framework that combined quantitative performance metrics and qualitative hierarchical-information analyses. Across the primary inpatient dataset (NIS 2019) and a 100-record MIMIC-IV ICU sample, HRF consistently achieved the best balance of accuracy, variance explained, and computational efficiency. HNN captured fine-grained group patterns but exhibited systematic negative bias and higher resource demands, whereas HMM provided coefficient-level interpretability yet under-performed in predictive accuracy.

These results offer practical guidance: tree-based models are recommended for general-purpose hierarchical prediction; HNNs are valuable when group differentiation outweighs calibration; HMMs remain useful when formal statistical inference is paramount. The persistent difficulty of all models at the hospital tier highlights an open research problem in modeling intermediate hierarchical effects.

Limitations include reliance on healthcare data and a small external-validation sample. Future work should test additional domains, larger external cohorts, and hybrid or fairness-aware extensions that combine the complementary strengths identified here.

Overall, the findings confirm that tree-based architectures currently deliver the most robust and interpretable solution for hierarchical healthcare modeling.

CRediT authorship contribution statement

Marzieh Amiri Shahbazi: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Nasibeh Azadeh-Fard:** Supervision, Validation, Writing – review & editing, Project administration, Resources.

Declaration of competing interest

The authors declare no competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Potential publication fees may be covered by the AWARE-AI NSF Research Traineeship Program; this support does not constitute a conflict of interest.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data availability

The authors do not have permission to share data.

References

Al-Rubaye, R. H. K., & Türkben, A. K. (2024). Using artificial intelligence to evaluating detection of cybersecurity threats in ad hoc networks. *Babylonian Journal of Networking*, *2024*, 45–56.

Asampana Asosega, K., Adebanji, A. O., Aidoo, E. N., & Owusu-Dabo, E. (2024). Application of hierarchical/multilevel models and quality of reporting (2010–2020): A systematic review. *The Scientific World Journal*, *2024*(1), Article 4658333.

Bafandeh, A., Bin-Karim, S., Baheri, A., & Vermillion, C. (2018). A comparative assessment of hierarchical control structures for spatiotemporally-varying systems, with application to airborne wind energy. *Control Engineering Practice*, *74*, 71–83.

Baheri, A. (2025). Multilevel constrained bandits: A hierarchical upper confidence bound approach with safety guarantees. *Mathematics*, *13*(1), 149.

Bojic, I., Chen, J., Chang, S. Y., Ong, Q. C., Joty, S., & Car, J. (2023). Hierarchical evaluation framework: Best practices for human evaluation. ArXiv Preprint, arXiv:2310.01917.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees–crc press. *Boca Raton, Florida*, *685*.

Chen, D., Huang, X., Sun, X., Ma, W., & Zhang, S. (2016). A comparison of hierarchical and non-hierarchical Bayesian approaches for fitting allometric larch (Larix spp.) biomass equations. *Forests*, *7*, 18, URL https://api.semanticscholar.org/CorpusID:17409203.

Corigliano, S., Rosato, F., Ortiz Dominguez, C., & Merlo, M. (2021). Clustering techniques for secondary substations siting. *Energies*, *14*(4), 1028.

de Resende, M. D. V., & Alves, R. S. (2020). Linear, generalized, hierarchical, Bayesian and random regression mixed models in genetics/genomics in plant breeding. *Functional Plant Breeding Journal*, *2*(2).

Dowding, I., & Haufe, S. (2018). Powerful statistical inference for nested data using sufficient summary statistics. *Frontiers in Human Neuroscience*, *12*, 103.

Ebnehoseini, Z., Tabesh, H., Jangi, M., Deldar, K., Mostafavi, S. M., & Tara, M. (2021). Investigating evaluation frameworks for electronic health record: A literature review. *Open Access Macedonian Journal of Medical Sciences*, URL https://api.semanticscholar.org/CorpusID:233473705.

Fan, W., Guo, Z. H., Bouguila, N., & Hou, W. (2021). Clustering-based online news topic detection and tracking through hierarchical Bayesian nonparametric models. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. URL https://api.semanticscholar.org/CorpusID:235792246.

Gelman, A. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Goldstein, H. (2011). *Multilevel Statistical Models*. John Wiley & Sons.

Guo, Y., Luo, Y., He, Z., Huang, J., & Chen, J. (2020). Hierarchical neural architecture search for single image super-resolution. *IEEE Signal Processing Letters*, *27*, 1255–1259.

Guo, Z., Renaux, C., Bühlmann, P., & Cai, T. (2021). Group inference in high dimensions with applications to hierarchical testing. *Electronic Journal of Statistics*, *15*(2), 6633–6676.

Harbord, R. M., Whiting, P., Sterne, J. A. C., Egger, M., Deeks, J. J., Shang, A., & Bachmann, L. M. (2008). An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology*, *61*(11), 1095–1103.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2). New York: Springer.

Heck, R. H., & Thomas, S. L. (2015). *An Introduction to Multilevel Modeling Techniques*. Routledge.

Hooda, R. (2017). Digital image processing through hierarchical clustering methods, tree classifier of data mining. In *Proceedings*. URL https://api.semanticscholar.org/CorpusID:212517138.

Hox, J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel Analysis: Techniques and Applications*. Routledge.

Johnson, A. E. W., Bulgarelli, L., Pollard, T. J., Horng, S., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV (version 2.2). http://dx.doi.org/10.13026/6mm1-ek67, https://physionet.org/content/mimiciv/2.2/.

Koltcov, S., Ignatenko, V., Terpilovskii, M., & Rosso, P. (2021). Analysis and tuning of hierarchical topic models based on Renyi entropy approach. *PeerJ Computer Science*, *7*, Article e608.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on Optimization*, *9*(1), 112–147.

Li, N., Yuan, R., & Zheng, S. (2023). Trade-offs between poverty alleviation and household energy intensity in China. *Environmental Impact Assessment Review*, *98*, Article 106957.

Lindstrom, M. J., & Bates, D. M. (1988). Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Ma, J., Yu, M. K., Fong, S., Ono, K., Sage, E., Demchak, B., Sharan, R., & Ideker, T. (2018). Using deep learning to model the hierarchical structure and function of a cell. *Nature Methods*, *15*(4), 290–298.

McNabb, C. B., & Murayama, K. (2021). Unnecessary reliance on multilevel modelling to analyse nested data in neuroscience: When a traditional summary-statistics approach suffices. *Current Research in Neurobiology*, *2*, Article 100024.

Moen, E. L., Fricano-Kugler, C. J., Luikart, B. W., & O'Malley, A. J. (2016). Analyzing clustered data: why and how to account for multiple observations nested within a study participant? *Plos One*, *11*(1), Article e0146721.

Munmun, F. A., & Khatun, S. (2022). A hybrid method: Hierarchical agglomerative clustering algorithm with classification techniques for effective heart disease prediction. *International Journal of Research and Innovation in Applied Science*, URL https://api.semanticscholar.org/CorpusID:251634998.

Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization* (2nd). Springer.

Opoku Larbi, S., Rangita, A., & Otieno, J. (2024). Advancing hierarchical model: Evaluating performance, interpretability and implications. *IJFMR*, *6*(6), http://dx.doi.org/10.36948/ijfmr.2024.v06i06.30796.

Raudenbush, S. W. (2002). *Advanced quantitative techniques in the social sciences series*, *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Ritzert, M., Turishcheva, P., Hansel, L., Wollenhaupt, P., Weis, M., & Ecker, A. (2025). Hierarchical clustering with maximum density paths and mixture models. arXiv preprint arXiv:2503.15582.

Salditt, M., Humberg, S., & Nestler, S. (2023). Gradient tree boosting for hierarchical data. *Multivariate Behavioral Research*, *58*(5), 911–937.

da Silva Goncalves, J., Manduchi, L., Vandenhirtz, M., & Vogt, J. E. (2024). Hierarchical clustering for conditional diffusion in image generation. ArXiv E-Prints, arXiv–2410.

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *36*(2), 111–133.

Tambe-Jagtap, S. N., & Jaaz, Z. A. (2023). Integrative genomic and proteomic profiling for personalized oncological treatments: enhancing therapeutic efficacy and reducing adverse effects in breast cancer patients. *SHIFAA*, *2023*, 1–9.

Tessler, J. M. (2014). *Three-level models for partially nested data structures*. University of California, Los Angeles.

Thrane, E., & Talbot, C. (2019). An introduction to Bayesian inference in gravitational-wave astronomy: Parameter estimation, model selection, and hierarchical models. *Publications of the Astronomical Society of Australia*, *36*, Article e010.

Vermunt, J. K., & Magidson, J. (2005). Hierarchical mixture models for nested data structures. In *Classification—the ubiquitous challenge: proceedings of the 28th annual conference of the gesellschaft für klassifikation eV university of dortmund, March 9–11, 2004* (pp. 240–247). Springer.

Virupakshappa, K., Marino, M., & Oruklu, E. (2018). A multi-resolution convolutional neural network architecture for ultrasonic flaw detection. In *2018 IEEE international ultrasonics symposium* (pp. 1–4). IEEE.

Wikle, C. K. (2016). Hierarchical models for uncertainty quantification: An overview. In *Handbook of uncertainty quantification* (pp. 193–218). Springer International Publishing Cham.

Wikle, C. K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*(2), 175–203.

Xiao, M., Wu, M., Qiao, Z., Fu, Y., Ning, Z., Du, Y., & Zhou, Y. (2023). Interdisciplinary fairness in imbalanced research proposal topic inference: A hierarchical transformer-based method with selective interpolation. *ACM Transactions on Knowledge Discovery from Data*.

Yuan, P., Lin, S., Cui, C., Du, Y., Guo, R., He, D., Ding, E., & Han, S. (2020). HS-ResNet: Hierarchical-split block on convolutional neural network. ArXiv Preprint, arXiv:2010.07621.

Zaidan, M. A., Harrison, R. F., Mills, A. R., & Fleming, P. J. (2015). Bayesian hierarchical models for aerospace gas turbine engine prognostics. *Expert Systems with Applications*, *42*(1), 539–553.

Zeng, W., Mukherjee, S., Caudillo, A., Forman, J., & Panzer, M. B. (2021). Evaluation and validation of thorax model responses: A hierarchical approach to achieve high biofidelity for thoracic musculoskeletal system. *Frontiers in Bioengineering and Biotechnology*, *9*, Article 712656.

Zhang, L., Cheng, L., Li, H., Gao, J., Yu, C., Domel, R., Yang, Y., Tang, S., & Liu, W. K. (2021). Hierarchical deep-learning neural networks: Finite elements and beyond. *Computational Mechanics*, *67*, 207–230.

Zhao, X., Du, J., & Wang, Z. (2023). HCS-R-HER: Hierarchical reinforcement learning based on cross subtasks rainbow hindsight experience replay. *Journal of Computational Science*, *72*, Article 102113.

Zyzanski, S. J., Flocke, S. A., & Dickinson, L. M. (2004). On the nature and analysis of clustered data. *The Annals of Family Medicine*, *2*(3), 199–200.