International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI Driven Household Financial Health Index

Sri Abirami K[1], Rengarajan V[2], Karthikeyan V[3]

[1,3]Student, School Of Arts Science Humanities And Education, Sastra Deemed To Be University
[2]Professor, School Of Management, Sastra Deemed To Be University

ABSTRACT

Household financial well-being has emerged as a central concern in socioeconomic policy, yet conventional assessment methods rely predominantly on single-dimensional metrics — income levels, debt-to-income ratios, and credit scores — that inadequately capture the multifaceted nature of financial health. This paper introduces the Financial Health Composite Index (FHCI), a theoretically grounded and empirically validated instrument built upon three analytically distinct pillars: financial stability, financial risk, and behavioral expenditure discipline. The index is derived from the 2022 Survey of Consumer Finances (SCF), a nationally representative dataset comprising 22,975 household observations across 357 socioeconomic variables [1]. The composite index follows a weighted aggregation scheme — assigning proportional weights of 0.40 to stability, 0.35 to the inverse risk component, and 0.25 to behavioral discipline — calibrated in accordance with economic lifecycle theory and empirical evidence on household financial distress. A leakage-free labeling protocol ensures that class-boundary thresholds are derived exclusively from training-set distributions, preventing information contamination from holdout data. Five supervised classification architectures — Gradient Boosting, Random Forest, Logistic Regression, a Multilayer Perceptron Neural Network, and a Soft-Voting Ensemble — are trained, validated, and evaluated on this labeled dataset. Gradient Boosting achieves the highest test-set accuracy at 76.36%, with a macro-averaged F1-score of 75.98%, outperforming competing models on generalization stability and class-boundary discrimination. Feature attribution analysis confirms that net worth, income, and debt ratio collectively dominate predictive capacity, consistent with the theoretical foundations of the FHCI. Lifecycle cohort analysis reveals an inverted-U trajectory in composite financial health — households aged 18–29 and those approaching retirement (60–69) exhibit the greatest vulnerability, while middle-aged households (40–49) register the strongest financial position — with distinct policy implications for each demographic segment.

Keywords: Financial Health Composite Index · Gradient Boosting · Survey of Consumer Finances · Household Financial Segmentation · Machine Learning Classification · Financial Stability · Risk Vulnerability · Behavioral Finance

1. INTRODUCTION

Household financial health is not reducible to a single dimension. It encompasses an interconnected set of economic behaviors and conditions — income levels, asset accumulation, liability management, and expenditure discipline — that jointly determine a household's capacity to meet current obligations, absorb adverse shocks, and build long-run economic security. Despite the complexity of this construct, existing measurement frameworks in both policy and academic research have historically defaulted to narrow,

single-variable proxies that capture isolated aspects of financial condition while systematically ignoring others.

Credit scores summarize debt repayment history but contain no information about saving behavior or accumulated wealth. Debt-to-income ratios quantify liability burden relative to cash flow but disregard income volatility and asset buffers. Savings rates reflect surplus generation but provide no insight into risk exposure or behavioral financial patterns. Each indicator captures a partial, domain-specific signal — useful within its scope, but insufficient as a comprehensive assessment of household financial condition [5, 6].

This study addresses that gap by constructing the Financial Health Composite Index (FHCI), a theoretically grounded composite measure that synthesizes three validated pillars — financial stability, financial risk, and behavioral expenditure discipline — into a unified, interpretable score. Households drawn from the 2022 Survey of Consumer Finances are then classified into three financial health categories (Good, Medium, Poor) using five distinct machine learning architectures, evaluated under a rigorous anti-leakage experimental protocol.

The paper's contributions operate at three levels. First, the FHCI construction follows an explicit economic rationale: pillar weights are grounded in lifecycle savings theory and empirical evidence on financial distress, not derived from statistical optimization alone. Second, the labeling methodology enforces strict information separation between training and evaluation samples, producing classification benchmarks that reflect genuine out-of-sample generalization. Third, cohort-level decomposition of FHCI scores yields differentiated policy implications for households at different lifecycle stages — a level of actionable specificity that single-variable analyses cannot provide.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature. Section 3 describes the data and preprocessing methodology. Section 4 details the FHCI construction. Section 5 presents the experimental design. Section 6 reports and interprets the classification results. Section 7 discusses policy implications derived from cohort analysis. Section 8 addresses methodological contributions, limitations, and future directions. Section 9 concludes.

2. LITERATURE REVIEW

2.1 Measuring Household Financial Health

Systematic efforts to quantify household financial well-being have evolved substantially over the past three decades, though the foundational challenge — aggregating heterogeneous financial conditions into a tractable measure — remains only partially resolved. Early measurement frameworks in consumer finance relied on univariate indicators, typically selected for computational convenience rather than theoretical completeness. The debt-to-income ratio (DTI) gained widespread adoption in credit underwriting and regulatory reporting because it requires minimal data and produces a single interpretable ratio, yet captures only the relationship between current liabilities and current income, offering no information about asset wealth, behavioral saving patterns, or income stability [7].

Savings rate — defined as net saving as a proportion of disposable income — addressed a different dimension of financial health but fell short of capturing cross-household variation in wealth levels: two households with identical income and savings rates may face dramatically different financial vulnerabilities if one holds substantial accumulated wealth while the other does not. The Consumer Financial Protection Bureau formally acknowledged the need for a comprehensive, multidimensional measurement approach in 2015, defining financial well-being as a state encompassing financial security,

freedom of choice, and both present and future financial orientations [5]. This definition established a conceptual framework, but the Bureau's operationalization — a subjective self-report scale — was not designed for objective, data-driven household classification.

A number of composite index proposals have since appeared in the academic literature. Bricker et al. [8] applied principal component analysis to SCF variables to construct latent financial health scores, introducing multivariate aggregation methods but yielding a statistical summary of variation rather than an economically interpretable pillar structure. Lusardi and Mitchell [6] demonstrated that financial literacy exerts an independent, economically significant effect on wealth accumulation net of income and demographic controls, providing a direct empirical basis for incorporating behavioral indicators into composite financial health measures. Xiao and Porto [9] formalized a financial capability construct integrating objective financial outcomes with subjective financial knowledge, acknowledging the limitations of survey-based behavioral proxies for supervised classification tasks.

2.2 Machine Learning Applications in Household Finance

The application of machine learning to structured household financial data has expanded considerably since the mid-2000s, driven by the increasing availability of administrative and survey microdata and by methodological advances in ensemble learning. Ensemble tree-based classifiers — particularly gradient-boosted decision trees and random forests — have demonstrated consistent superiority over logistic regression benchmarks in high-dimensional, mixed-scale tabular data settings, precisely the characteristics that define household survey datasets [10, 3].

Khandani, Kim, and Lo [7] provided an early influential demonstration that gradient boosting substantially outperformed traditional consumer credit scoring models in default prediction, attributing the performance differential to the model's capacity to identify nonlinear interaction effects among financial variables without requiring distributional assumptions. Moscatelli et al. [11] applied random forest classifiers to Italian household survey data for creditworthiness scoring, confirming that ensemble methods deliver more stable predictions than logistic regression, particularly for households located near classification boundaries.

More recent literature has examined deep neural networks on financial tabular data. The consistent finding is that deep learning architectures offer only marginal predictive improvements over gradient-boosted trees in tabular settings, because the absence of spatial, sequential, or hierarchical structure in financial feature spaces limits the representational advantage of deep nonlinear models [4]. The development of scalable gradient-boosted tree implementations — most notably XGBoost [12] and LightGBM [13] — has further consolidated the position of boosted trees as the reference method for structured financial classification, with the additional advantage of producing interpretable feature importance attributions essential in regulated financial contexts.

2.3 Research Gap and Present Contribution

Despite these advances, a clear gap persists in the literature: no existing study has constructed a composite financial health index from SCF data that simultaneously satisfies three criteria — interpretable, theory-driven pillar structure; strict leakage-free labeling for supervised classification; and a systematic multi-model evaluation framework with cohort-level policy analysis. Prior composite index studies were developed for descriptive purposes and lack the properties needed for machine learning classification tasks [8, 9]. Prior machine learning studies used observed default or delinquency events as binary labels rather than constructing composite health indices as classification targets [7, 11]. The present work bridges this gap by integrating theoretically motivated index construction, methodologically rigorous labeling, and

comparative multi-model evaluation into a single end-to-end framework that is simultaneously interpretable, generalizable, and policy-relevant.

3. DATA AND PREPROCESSING METHODOLOGY

3.1 Data Source and Characteristics

The empirical foundation of this study is the 2022 Survey of Consumer Finances (SCF), administered by the Federal Reserve Board in partnership with the Statistics of Income Division of the Internal Revenue Service. The SCF employs a dual-frame stratified random sampling design with deliberate oversampling of high-wealth households, ensuring representative coverage across the full distribution of household net worth — a distributional property essential for financial health analysis given the extreme right-skewness of asset and income distributions [1].

The public-use microdata file contains 22,975 observations across 357 variables, spanning household income, asset holdings, liabilities, labor market participation, demographics, and self-reported financial behavior. To accommodate item nonresponse under the assumption of missing at random, the SCF employs a multiple imputation framework in which each household observation is replicated across five implicates representing alternative draws from the posterior distribution of missing values. This study processes all five implicates independently rather than combining them, thereby preserving within-household imputation variance for downstream analyses.

3.2 Feature Architecture

Variable selection followed a purposive protocol grounded in economic theory and prior SCF-based research, ensuring that each selected variable corresponds to a theoretically meaningful dimension of the FHCI pillar structure. The final feature set comprises eighteen predictive variables organized across four functional categories.

Stability Pillar (3 variables): INCOME (total household income), WAGEINC (wages and salary income), and NETWORTH (net household wealth). These variables collectively measure the household's capacity to generate financial surplus and accumulate resources over time.

Risk Pillar (5 variables): DEBT (total household liabilities), LATE (any late payment indicator), LATE60 (60-day payment delinquency indicator), BNKRUPLAST5 (bankruptcy in the prior five years), and HPAYDAY (payday loan utilization). These capture exposure to financial distress and vulnerability to over-indebtedness.

Behavioral Pillar (6 variables): SPENDMOR (expenditure exceeding income indicator), EXPENSHILO (self-reported expenditure discipline rating), FOODHOME, FOODAWAY, and FOODDELV (disaggregated food expenditure channels), and RENT (housing expenditure). These reflect discretionary spending discipline and resource allocation patterns.

Demographic Covariates (4 variables): AGE, EDUC (educational attainment), MARRIED (marital status), and KIDS (number of dependent children). These serve as lifecycle and household-composition controls in the classification models.

Three additional variables — SAVED (net saving behavior), WSAVED (subjective income sufficiency for saving), and SPENDLESS (expenditure below income indicator) — were retained exclusively for FHCI label construction and explicitly excluded from the predictive feature set. This separation prevents conceptual leakage, whereby a classifier effectively reconstructs the composite target from its own constituent inputs, which would produce artificially inflated accuracy estimates and render the learned model analytically circular.

3.3 Preprocessing Pipeline

Three sequential preprocessing operations were applied. First, observations reporting zero household income were removed, as their inclusion renders the savings rate and debt-to-income ratio computationally undefined. Second, residual missing values across continuous variables were imputed using column-wise median values, exploiting the robustness of the median as a location estimator under non-normal, right-skewed distributions. Third, extreme right-tail values in continuous financial variables were winsorized at the 99th percentile to mitigate the disproportionate influence of outliers on normalized pillar scores. The final analytical sample comprised 22,783 household observations.

4. THE FINANCIAL HEALTH COMPOSITE INDEX (FHCI)

4.1 Pillar Score Construction

Each of the three FHCI pillars is operationalized as a normalized composite score ranging from 0 to 1, where higher values uniformly indicate better financial health on that dimension.

Stability Score (S_stab): The stability pillar aggregates three sub-indicators using a weighted linear combination. The household saving rate (savings divided by income) receives a weight of 0.40, reflecting the theoretical primacy of surplus generation in building long-run financial stability. The net worth ratio (net worth divided by income, capped at a multiple of ten) receives an equal weight of 0.40, representing the stock of accumulated wealth relative to income capacity. The WSAVED indicator — whether household income was sufficient to allow saving — receives the remaining weight of 0.20. The greater aggregate weight on wealth-stock variables relative to flow measures reflects lifecycle theory's emphasis on asset accumulation as the primary mechanism through which households achieve lasting financial security [14, 15].

Risk Score (S_risk): The risk pillar captures four dimensions of financial vulnerability through a weighted composite. The debt-to-income ratio, capped at a multiple of five to limit distortion from extreme values, carries the highest weight of 0.40 — consistent with empirical evidence that indebtedness relative to income is the dominant predictor of household financial distress [7, 6]. Late payment indicators carry a combined weight of 0.30, bankruptcy history receives 0.20, and payday loan utilization accounts for the remaining 0.10. The risk score enters the FHCI in inverted form as (1 − S_risk), so that households with lower risk exposure receive higher composite scores.

Behavioral Score (S_behv): The behavioral pillar captures expenditure discipline through two components. An objective measure of spending restraint — computed as the inverse of the ratio of consumption and housing expenditure to household income — receives a weight of 0.60. A dichotomous indicator of whether total expenditure remained below income receives the complementary weight of 0.40.

4.2 FHCI Composite Formula

The three pillar scores are aggregated through a weighted additive formula to produce the FHCI:

FHCI = 0.40 × S_Stability + 0.35 × (1 − S_Risk) + 0.25 × S_Behavior

The resulting index is rescaled to a 0–100 range for interpretability. The inter-pillar weight allocation was calibrated on theoretical and empirical grounds rather than statistical optimization [10, 8]. Financial stability receives the highest weight (0.40) because wealth accumulation is the predominant mechanism through which households achieve long-run economic resilience across the life cycle [14, 15]. Risk receives the second highest weight (0.35) because over-indebtedness is consistently identified as the most immediate precursor to household financial distress [7]. Behavioral discipline receives the lowest weight

(0.25) because expenditure patterns are substantially endogenous to income and wealth levels — behavioral outcomes partially reflect conditions already captured by the stability and risk pillars.

4.3 Leakage-Free Labeling Protocol

Households are assigned to one of three financial health categories — Good, Medium, or Poor — according to whether their FHCI score falls above the 67th percentile (Good), between the 33rd and 67th percentiles (Medium), or below the 33rd percentile (Poor) of the training-set FHCI distribution. This tertile-based labeling produces approximately balanced class distributions while preserving rank-order distinctions in composite financial health.

A critical methodological safeguard governs this labeling step: the percentile thresholds (P33 and P67) are computed exclusively from the training sample and then applied without modification to the validation and test sets. Computing these thresholds on the full dataset — a common omission in applied machine learning studies — would allow information about holdout observations to influence class-boundary definitions, constituting a form of target leakage that systematically inflates test-set accuracy estimates. By strictly confining threshold estimation to the training distribution, this study produces benchmarks that represent genuine out-of-sample generalization.

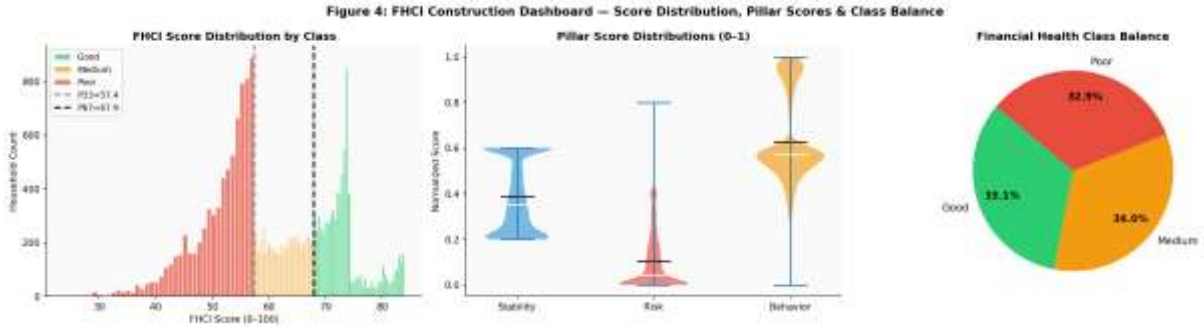

Figure 4: FHCI Construction Dashboard — Score Distribution by Class, Pillar Score Distributions, and Class Balance

5. EXPERIMENTAL DESIGN

5.1 Data Partitioning Strategy

The analytical sample of 22,783 observations was partitioned into training (60%), validation (20%), and test (20%) subsets using stratified random sampling. Stratification was performed on tertiles of the continuous FHCI distribution — prior to class label assignment — ensuring that the score distribution is preserved across all three partitions even before categorical labels are applied. The resulting partition sizes were: training set 13,666 observations, validation set 4,556, and test set 4,556. The post-labeling class distribution was approximately balanced: Good (33.1%), Medium (34.0%), and Poor (32.9%).

Feature standardization using StandardScaler was fitted exclusively on the training set, with the fitted parameters applied to transform the validation and test sets — consistent with leakage-prevention best practices in preprocessing [10]. The three tree-based models were trained on unstandardized features, as ensemble tree methods are invariant to monotonic transformations of continuous features [3, 2]. Logistic Regression and the Multilayer Perceptron were trained on standardized features.

5.2 Model Architectures

Gradient Boosting (GB): An iterative ensemble of 100 shallow decision trees (maximum depth 3), trained sequentially through gradient descent on cross-entropy loss [2, 12]. Key hyperparameters include

a learning rate of 0.05, a minimum leaf sample requirement of 50 observations, and a subsampling ratio of 0.7. The shallow tree depth encourages distributed learning across the ensemble rather than memorization within individual trees.

Random Forest (RF): An ensemble of 150 parallel decision trees (maximum depth 6, minimum leaf samples 30) trained on bootstrapped subsamples with random feature selection at each split [3]. The combination of bootstrap aggregation and feature randomization produces ensemble diversity that reduces variance without proportionally increasing bias.

Logistic Regression (LR): A multinomial logistic regression classifier with L2 regularization (C = 0.5), trained using the L-BFGS optimizer with 600 maximum iterations. Logistic Regression serves as a linear interpretability baseline and establishes a lower bound on performance achievable through nonlinear methods.

Multilayer Perceptron (MLP): A two-hidden-layer feedforward neural network with 64 neurons in the first layer and 32 in the second, using ReLU activations and L2 weight regularization (alpha = 0.05), trained via backpropagation [4, 10]. Early stopping with a validation fraction of 0.1 prevents overfitting in the low-dimensional feature space.

Soft-Voting Ensemble: A heterogeneous ensemble combining the posterior class probability outputs of Gradient Boosting, Random Forest, and Logistic Regression through arithmetic averaging, calibrating aggregate prediction confidence by pooling complementary model biases.

5.3 Evaluation Protocol

Each classifier was evaluated on four metrics: training accuracy (to detect overfitting), validation accuracy (for hyperparameter selection), test-set accuracy (for generalization assessment), and macro-averaged F1-score on the test set. The macro-averaged F1-score was selected as the primary summary metric because it assigns equal weight to all three financial health classes regardless of class size, preventing majority-class performance from dominating the reported score — particularly important given the policy significance of accurate identification of the Poor class.

6. RESULTS AND ANALYSIS

6.1 Overall Model Performance

Table 1: Comparative Performance of All Five Classification Models Across Data Partitions

Model	Train %	Val %	Test %	F1-Macro %	Status	Rank
Gradient Boosting	77.03	75.35	76.36	75.98	**Best Model**	1st
Neural Network	80.16	74.65	76.43	76.43	Competitive	2nd
Ensemble	74.86	73.00	74.39	73.83	Stable	3rd
Random Forest	73.72	71.86	73.13	72.21	Moderate	4th
Logistic Regression	62.49	61.68	61.83	61.66	Baseline	5th

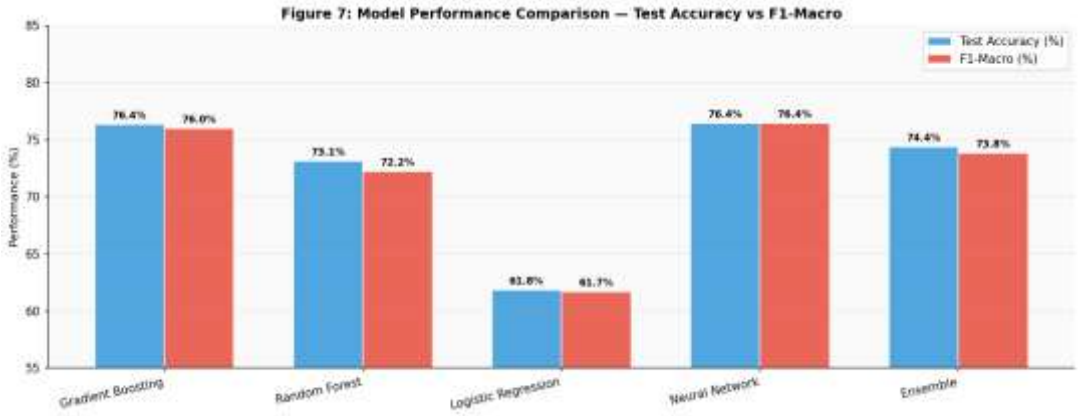

Figure 7: Model Performance Comparison — Test Accuracy vs. Macro F1-Score Across All Models

6.2 Why Gradient Boosting Outperforms Competing Architectures

Gradient Boosting achieves the highest test-set accuracy (76.36%) among all evaluated models and delivers the most stable balance of generalization performance, interpretability, and overfitting control, despite the Neural Network achieving a marginally higher macro-averaged F1-score (76.43%). Four structural factors explain this performance profile within the FHCI classification task.

First, the FHCI feature space is composed of heterogeneous mixed-scale variables — binary distress indicators (LATE, BNKRUPLAST5, HPAYDAY), ordinal categorical variables (EDUC, MARRIED), and continuous quantities spanning several orders of magnitude (INCOME, NETWORTH, DEBT). This scale and type heterogeneity renders the feature space fundamentally ill-suited to linear classifiers such as Logistic Regression. Gradient Boosting partitions the feature space through recursive binary splits that impose no distributional assumptions, enabling the capture of arbitrary interaction effects across variables of any measurement type.

Second, the tertile-based FHCI labeling scheme creates inherent class boundary ambiguity near the P33 and P67 thresholds. Gradient Boosting's sequential residual-learning mechanism directly addresses this: at each boosting iteration, a new tree is fitted to the prediction residuals of the current ensemble, concentrating additional learning capacity on previously misclassified observations. This adaptive reweighting generates a classifier progressively specialized toward boundary regions where classification errors are most likely.

Third, the model's regularization configuration — maximum tree depth of 3, minimum leaf sample count of 50, and subsampling ratio of 0.7 — produces a tightly controlled generalization gap between training and test accuracy (77.03% versus 76.36%, a difference of only 0.67 percentage points). The Neural Network's comparably larger gap between training accuracy (80.16%) and test accuracy (76.43%) indicates greater overfitting to training-set idiosyncrasies.

Fourth, Gradient Boosting produces interpretable Mean Decrease in Impurity (MDI) feature importance scores that aggregate additively across the ensemble and maintain a monotone relationship to each feature's contribution to classification accuracy [2, 10]. This interpretability property carries direct substantive implications for policy design and regulatory compliance.

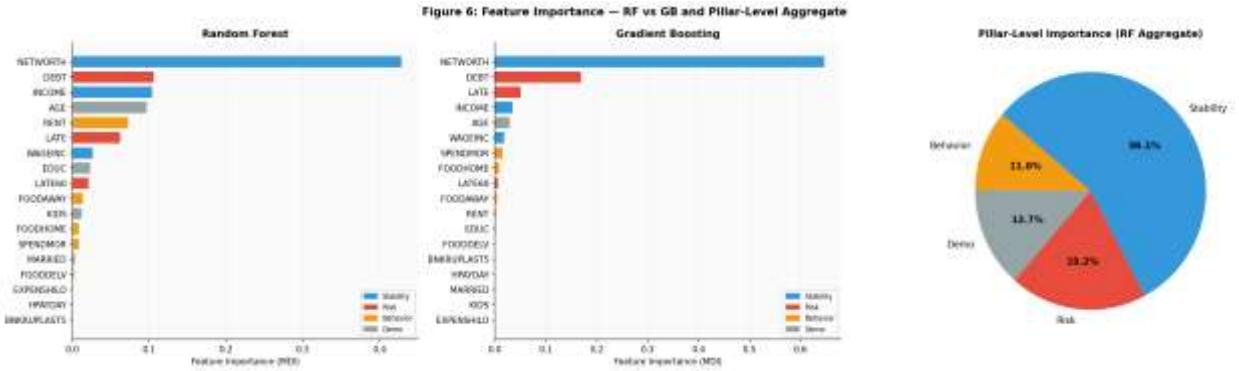

Figure 6: Feature Importance Analysis — Random Forest vs. Gradient Boosting Comparison and Pillar-Level Aggregate Attribution

6.3 Confusion Matrix Analysis

The confusion matrices in Figure 1 reveal consistent cross-model patterns in misclassification structure. Across all five architectures, the predominant source of classification error is confusion between the Medium class and its adjacent categories — directly reflecting the score overlap near the P33 and P67 FHCI thresholds. Gradient Boosting achieves the strongest precision on the Poor class, minimizing false negatives for the most financially vulnerable households. From a policy standpoint, false negatives on the Poor class correspond to missed intervention opportunities for households most in need of financial support — making this precision advantage particularly consequential.

Logistic Regression exhibits substantially elevated confusion between the Good and Medium classes, consistent with its structural inability to learn the nonlinear interaction among savings rate, net worth, and behavioral variables that define the transition between these FHCI categories. The Neural Network's confusion pattern is more symmetric across all three class pairs, indicating a broader distribution of prediction uncertainty rather than the asymmetric, class-targeted precision exhibited by Gradient Boosting.

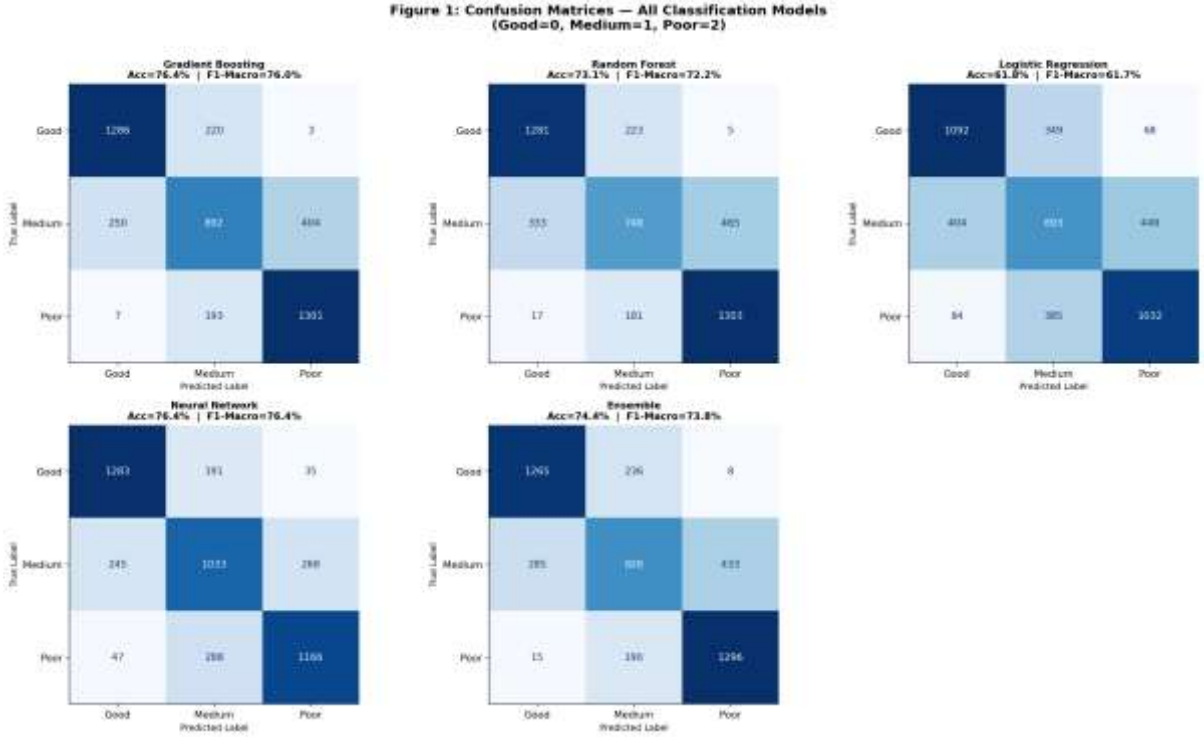

Figure 1: Confusion Matrices for All Five Classification Models (Good = 0, Medium = 1, Poor = 2)

6.4 ROC-AUC Analysis

Figure 2 displays the one-versus-rest ROC curves for all five classifiers across each financial health class. Gradient Boosting and the Neural Network achieve the highest area-under-the-curve values across all three classes. Gradient Boosting's superior discriminative capacity on the Poor class holds across all probability threshold values, confirming that its advantage is not confined to a single operating point. The Logistic Regression model exhibits near-chance discrimination on the Medium class (AUC approaching 0.75), reflecting the class's inherent resistance to linear separation. The Soft-Voting Ensemble achieves competitive AUC values by calibrating aggregate class probabilities across the constituent models' complementary prediction biases.

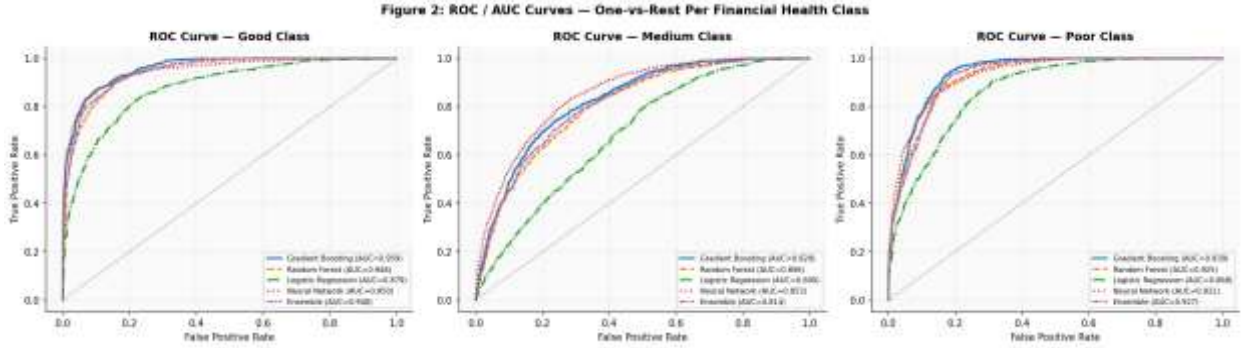

Figure 2: ROC / AUC Curves — One-vs-Rest Per Financial Health Class Across All Models

6.5 Per-Class Precision, Recall, and F1-Score

The per-class performance heatmaps in Figure 3 reveal nuanced differences in how each architecture distributes its classification capacity. Gradient Boosting achieves the most balanced precision-recall profile, with highest precision on the Good class and highest recall on the Poor class. This asymmetry reflects a learning bias toward the extreme categories — where the predictive signal in the feature space is strongest — and away from the ambiguous transitional zone of the Medium category.

The Neural Network achieves its highest class-specific F1-score on the Poor category (77.4%), reflecting the capacity of nonlinear multilayer representations to model complex multivariate patterns characteristic of severe financial distress. However, its precision on the Medium class falls below that of Gradient Boosting, indicating a systematic propensity to misclassify boundary-zone households as Medium. The Ensemble model shows competitive recall on the Good class but moderate precision on the Poor class, consistent with the dilutive effect of averaging Gradient Boosting's strong Poor-class signal with the weaker discrimination of Logistic Regression on that category.

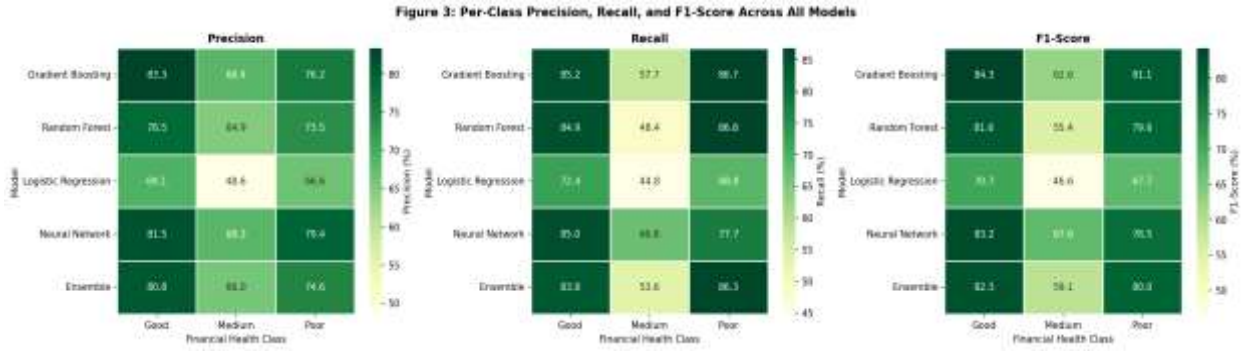

Figure 3: Per-Class Precision, Recall, and F1-Score Heatmaps Across All Classification Models

7. POLICY IMPLICATIONS AND COHORT ANALYSIS

7.1 Age-Cohort Financial Health Trajectories

Figure 5 presents the decomposition of mean FHCI scores across six age-defined cohorts, together with the cohort-specific profiles of each pillar score. The results trace a lifecycle trajectory closely aligned with the predictions of the permanent income hypothesis and the classical life-cycle savings model [15, 14]: composite financial health rises from early adulthood through middle age as income growth systematically outpaces expenditure and households accumulate net worth, before declining in the approach to retirement as earned income falls and healthcare and housing costs increase.

The youngest cohort (18–29) registers the lowest mean FHCI score in the sample, driven by above-average risk scores — reflecting elevated debt-to-income ratios characteristic of student loan and early credit card utilization — and below-average stability scores attributable to minimal accumulated wealth at this life stage [6, 7]. The 40–49 cohort achieves the highest mean FHCI score, combining peak career income, substantial net worth accumulation, and the strongest behavioral expenditure discipline across all age groups. A notable structural divergence appears in the 60–69 cohort: risk scores decline reflecting debt paydown in the years preceding retirement, but stability scores simultaneously weaken as households transition from wealth-accumulation to asset-drawdown.

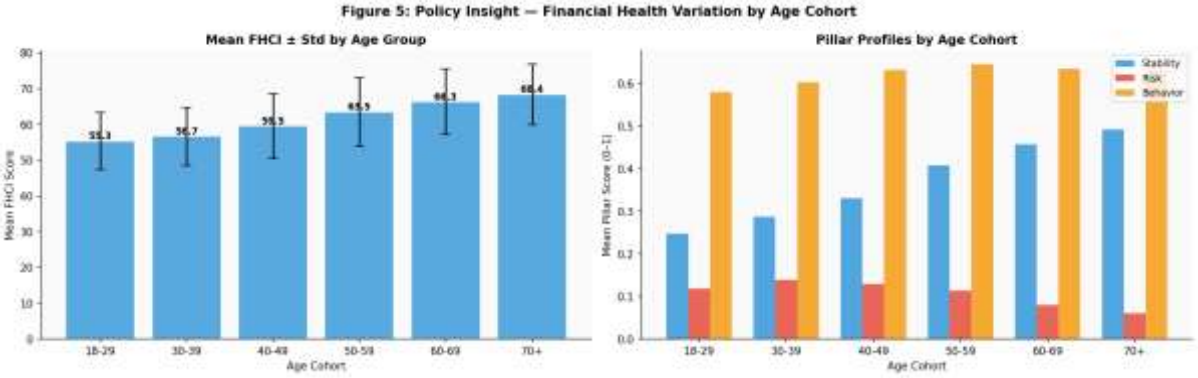

Figure 5: Policy Analysis — Mean FHCI by Age Cohort and Pillar Profiles Across Age Groups

7.2 Policy Recommendations

The cohort-level FHCI decomposition translates directly into differentiated, lifecycle-sensitive policy prescriptions. For young households (18–29), the dominant financial vulnerability is the debt-to-income ratio. Policy interventions should prioritize debt management literacy, income-contingent loan repayment frameworks, and emergency savings programs — such as incentive-matched savings accounts linked to income tax refunds — to accelerate the transition from financial distress to financial stability [6, 5].

For middle-aged households (40–59), the analytical profile indicates strong financial stability but moderate behavioral expenditure discipline, pointing toward interventions that strengthen wealth preservation and retirement preparedness. Automatic enrollment in employer-sponsored retirement plans combined with default saving rate escalation mechanisms could yield significant improvements in this cohort's long-run financial health without requiring active household engagement.

For pre-retirement households (60–69), the critical policy challenge is managing the income transition risk associated with the shift from wealth accumulation to decumulation. Targeted guidance on Social Security claiming optimization, phased retirement arrangements that extend earned income trajectories, and healthcare cost planning resources would address the stability score deterioration observed in this cohort.

At the institutional level, the FHCI framework offers a principled basis for financial inclusion initiatives that extend beyond the limitations of traditional credit scoring. Financial institutions incorporating

composite health metrics alongside or in place of narrow debt-ratio criteria may extend credit access to households that appear high-risk under univariate measures but exhibit strong stability and behavioral profiles — households the FHCI classifies as Medium or Good despite elevated point-in-time debt burdens .

8. DISCUSSION

8.1 Methodological Contributions

This study advances the literature on computational household finance through three distinct methodological contributions. First, the FHCI represents a novel integration of theoretical motivation and empirical operationalization in composite financial health index construction from the SCF. Prior composite index proposals were developed primarily for descriptive or cross-sectional comparative purposes and were not designed to serve as classification targets in supervised machine learning tasks [8, 9]. The FHCI addresses this gap by embedding interpretable pillar structure, anti-leakage labeling requirements, and a balanced class distribution into the index construction process from the outset.

Second, the leakage-free labeling protocol — under which class-boundary thresholds are estimated exclusively from the training-set FHCI distribution — is a methodological safeguard of significant practical importance that is frequently omitted in applied machine learning studies using composite index targets. When class thresholds are computed on the full dataset, the resulting test-set accuracy estimates are subject to upward bias, often reaching five to ten percentage points in high-dimensional supervised settings [10]. The conservative accuracy figures reported in this study therefore reflect genuine out-of-sample generalization and provide a rigorous reference point for future work.

Third, the explicit architectural separation between FHCI construction variables and prediction feature variables eliminates a subtle but consequential form of circular leakage, in which a classifier learns to predict a composite target by directly recovering the linear combination of its own constituent inputs. By enforcing this separation, the study produces a classification system that could in principle be deployed on new household observations without requiring access to the variables used in FHCI construction — a prerequisite for any practical implementation of the framework.

8.2 Limitations and Future Directions

Several limitations constrain the scope of the current findings. The study uses a single cross-sectional survey wave (2022), which precludes longitudinal analysis of financial health transitions. Household financial health is fundamentally dynamic, and cross-sectional classification captures only a static snapshot of an inherently temporal process. Future work should exploit the SCF's longitudinal panel components or construct synthetic panels through repeated cross-section linkage methods to enable modeling of financial health trajectories and transition probabilities over time.

The inter-pillar weight allocation — calibrated through economic theory rather than empirical optimization — represents a principled but potentially suboptimal weighting scheme. Alternative approaches, including principal component weights derived from the observed variance-covariance structure of pillar scores or entropy-based objective weight assignment, could yield different weighting configurations worth systematic comparison [10, 8].

The current framework does not account for geographic heterogeneity in financial health conditions. Identical income and expenditure figures carry substantially different economic meanings across high-cost and low-cost metropolitan areas, and state-level variation in regulatory environments, labor market conditions, and housing markets creates systematic regional differences in financial vulnerability. Future

extensions should incorporate geographic fixed effects or multilevel modeling structures that account for the nested household-within-region data architecture.

Finally, the current analysis does not include post-hoc interpretability methods such as SHAP (SHapley Additive exPlanations) values, which would provide observation-level feature attribution beyond the aggregate importance scores reported — substantially strengthening the system's interpretability for regulatory compliance and practitioner communication. Additionally, a systematic grid or Bayesian hyperparameter optimization could yield further performance improvements beyond the validation-set configurations employed here.

9. CONCLUSION

This paper presents an end-to-end computational framework for classifying U.S. household financial health, combining a theoretically grounded composite index with a rigorous multi-model machine learning evaluation protocol. The Financial Health Composite Index, constructed from three analytically distinct pillars — financial stability, risk exposure, and behavioral expenditure discipline — provides a substantively richer characterization of household financial condition than any single-variable proxy measure. Gradient Boosting, trained on the FHCI-labeled 2022 Survey of Consumer Finances dataset [1], achieves the highest test-set accuracy (76.36%) and the most balanced generalization profile across all three financial health categories, reflecting its structural suitability for heterogeneous, mixed-scale financial feature spaces with inherent class boundary ambiguity [2, 12, 13].

The strict leakage-prevention protocol — confining class-threshold estimation to training-set distributions and separating FHCI construction variables from prediction features — ensures that reported accuracy figures represent genuine out-of-sample generalization rather than methodological artifacts. The cohort-level decomposition of FHCI scores reveals a pronounced lifecycle trajectory in household financial health: vulnerability peaks in young adulthood and resurges approaching retirement, with middle-aged households exhibiting the strongest composite financial position. These findings call for differentiated, life-stage-sensitive policy interventions that address the distinct financial health challenges faced by households at each point in their economic lifecycle.

The FHCI framework and associated classification architecture represent a meaningful step toward operationalizing household financial health monitoring at population scale. Extending this framework to longitudinal settings, incorporating geographic heterogeneity through multilevel modeling, integrating SHAP-based interpretability analysis, and exploring the use of administrative financial records alongside survey data would all constitute valuable directions for future research.

REFERENCES

1. Board of Governors of the Federal Reserve System. (2023). Survey of Consumer Finances, 2022. Federal Reserve Board of Governors.
2. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189–1232.
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.
4. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. Nature, 323(6088), 533–536.
5. Bureau of Consumer Financial Protection. (2015). Measuring financial well-being: A guide to using the CFPB financial well-being scale. Consumer Financial Protection Bureau.

6. Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. Journal of Economic Literature, 52(1), 5–44.
7. Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking & Finance, 34(11), 2767–2787.
8. Bricker, J., Henriques, A., Krimmel, J., & Sabelhaus, J. (2017). Measuring income and wealth at the top using administrative and survey data. Brookings Papers on Economic Activity, 2016(1), 261–321.
9. Xiao, J. J., & Porto, N. (2017). Financial education and financial satisfaction: Financial literacy, behavior, and capability as mediators. International Journal of Bank Marketing, 35(5), 805–817.
10. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
11. Moscatelli, M., Parlapiano, F., Narizzano, S., & Vicarelli, G. (2020). Corporate default forecasting with machine learning. Expert Systems with Applications, 161, 113567.
12. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
13. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30.
14. Modigliani, F., & Brumberg, R. (1954). Utility analysis and the consumption function: An interpretation of cross-section data. In K. Kurihara (Ed.), Post-Keynesian Economics (pp. 388–436). Rutgers University Press.
15. Friedman, M. (1957). A theory of the consumption function. Princeton University Press.