Feature Selection and Reduction in Happiness Index Analysis: A Systematic Literature Review

Dani Ferdinan
Informatics engineering
vocational school
Universitas Logistik dan Bisnis Internasional
Bandung, Indonesia
daniferdinandall@gmail.com

Nisa Hanum Harani
Informatics engineering
vocational school
Universitas Logistik dan Bisnis Internasional
Bandung, Indonesia
nisa@ulbi.ac.id

Abstract—This study investigates the role and effectiveness of feature selection and feature reduction techniques in improving the accuracy, validity, and efficiency of predictive models for survey-based happiness indices. A Systematic Literature Review (SLR) was conducted following the PRISMA 2020 protocol, evaluating 40 peer-reviewed articles published between 2020 and 2025. The results demonstrate that feature selection methods namely wrapper, filter, and embedded approaches can significantly enhance model performance, yielding higher coefficients of determination (R^2) and lower prediction errors. Furthermore, the identification of relevant features has been shown to improve construct validity and the reliability of happiness indicators. The integration of feature selection and feature reduction techniques also contributes to more efficient and stable models, particularly in high-dimensional data contexts. However, the limited number of studies directly addressing happiness and the methodological heterogeneity across works pose challenges to the generalizability of the findings. This review provides valuable insights for establishing evidence-based practices and guiding strategic developments in future happiness index analytics.

Keywords— feature selection, feature reduction, happiness-index, machine learning, survey data, systematic literature review.

I. INTRODUCTION

Happiness indices have emerged as important indicators for measuring societal well-being, complementing traditional economic metrics such as Gross Domestic Product (GDP). In recent years, survey data have been widely employed to assess both individual and collective happiness, encompassing psychological, social, economic, and environmental dimensions [1]. However, the complexity and high dimensionality of happiness survey data present unique challenges in the analytical process, particularly in selecting relevant variables that significantly impact predictive outcomes [2].

In this context, feature selection methods play a crucial role in identifying the most influential attributes from survey data that contribute to happiness prediction. These techniques aim to reduce model complexity, improve accuracy, and prevent overfitting by filtering out irrelevant features [3]. In addition, feature reduction approaches such as Principal Component Analysis (PCA) are used to reduce data dimensionality while preserving essential information, thereby enabling more efficient and robust analyses [4]. Furthermore, Natural Language Processing (NLP) approaches have also been increasingly adopted in happiness-related studies, particularly when dealing with unstructured text such as open-ended survey responses, as exemplified by text-based sentiment analysis on social media platforms [5].

Several feature selection techniques, including filter, wrapper, and embedded methods, have been employed across various domains ranging from social sciences to biomedical and image processing applications [6]. Although these techniques have demonstrated effectiveness in improving model performance, studies that specifically examine their utility in the context of happiness surveys remain limited. Moreover, the relationship between feature selection outcomes and the construct validity and reliability of happiness indicators has often not been thoroughly analyzed [7].

In light of this background, the present systematic literature review is designed to address critical research questions fundamental to advancing the application of machine learning in happiness measurement, while establishing evidence-based practices for the analysis of social survey data. Based on a comprehensive analysis of 40 peer-reviewed scientific articles published between 2020 and 2025, this review focuses on the application of feature selection and feature reduction techniques to enhance both the performance and interpretability of models predicting happiness indices from survey-based data.

This review is guided by the following three primary research questions:

- RQ1: How do feature selection methods affect the accuracy of happiness index measurement using survey data?
- RQ2: To what extent do feature selection methods improve the validity and reliability of variables in survey-based happiness measurement?
- RQ3: What is the distribution of feature handling methods, and which is the most dominant?

Through a systematic synthesis of recent literature, this review aims to provide comprehensive and evidence-based insights into the role of feature selection and reduction methods in survey-based happiness measurement. It also seeks to establish methodological benchmarks, identify best practices, and offer strategic directions for future research at the intersection of data science, psychology, and social policy.

A. Feature Selection and Feature Reduction

In the analysis of complex and high-dimensional data, such as happiness survey data, feature processing becomes a critical aspect that influences the performance of predictive models. Feature selection is the process of selecting the most relevant subset of features from the entire set of available variables, aiming to improve model accuracy, reduce training time, and minimize overfitting. A study by Aouragh et al. (2024) integrated feature selection methods and dimension reduction techniques such as PCA, demonstrating significant performance improvements in disease diagnosis based on machine learning [8]. Feature selection methods are divided into three main categories: filter methods, wrapper methods, and embedded methods. For instance, the genetic algorithm approach used by Andrews et al. (2023) proved effective in significantly reducing prediction errors in the context of spectroscopy, highlighting the importance of feature selection as a crucial part of model optimization [9]. Feature selection techniques are generally categorized into three main types: filter, wrapper, and embedded methods [10].

1) Filter method: The filter method selects features based on their intrinsic statistical properties, independent of the machine learning algorithm that will be used. This approach utilizes metrics like Pearson's, Spearman's, or Kendall's correlation to evaluate and rank features according to their relationship with the target variable, after which a user-defined threshold is applied to select the most relevant ones [11]. The main advantage of the filter method is its low computational cost, as the selection is performed only once before model training. However, its primary weakness is that it ignores the interaction between the selected features and the learning algorithm, which can result in a suboptimal feature subset and poor model performance [12].

2) Wrapper method: The wrapper model utilizes a machine learning algorithm as an integral part of the feature evaluation function, where the performance of the classifier is used to score the usefulness of various feature subsets. Based on the inference calculated from the trained model, features are iteratively added or removed from the subset in a process that, while potentially having a massive time complexity, aims to find the most useful features [13]. Common examples of this method include Recursive Feature Elimination (RFE) and Backward Feature Elimination (BFE). While wrapper methods typically lead to better model performance than filter methods because they account for feature interactions and model bias, they are significantly more computationally expensive and have a higher risk of overfitting, especially when dealing with limited datasets [14].
3) Embedded methods: Embedded methods represent a class of feature selection techniques wherein the process of selecting an optimal feature subset is intrinsically integrated into the model training algorithm. These techniques leverage the model's internal mechanisms, such as regularization penalties, to evaluate and weight the contribution of each feature during the learning process. A primary example of this method is LASSO (Least Absolute Shrinkage and Selection Operator), which employs L1 regularization to shrink the coefficients of irrelevant features to exactly zero, thereby effectively performing feature elimination[15]. Other popular approaches include tree-based algorithms such as Random Forest and Gradient Boosting, which compute feature importance scores as a byproduct of the model-building process. By coupling feature selection with model training, embedded methods offer an efficient trade-off between computational performance and model accuracy, making them superior to model-agnostic filter methods and faster than computationally expensive wrapper methods [14].

Meanwhile, feature reduction aims to transform high-dimensional data representation into a lower dimension while retaining essential information. Techniques such as PCA or LDA transform the original features into new, more compact, and non-redundant feature [16]. In certain cases, semi-parametric approaches, such as those developed by Liu et al. (2023), enable robust feature selection in the presence of non-normal error distributions, yielding consistent feature selection even under skewed and heavy-tailed distributions [17].

The integration of feature selection and feature reduction has proven to result in more efficient and effective models, particularly in the context of complex and heterogeneous survey data, such as happiness index measurement.

B. Concept of Happiness Index and Measurement from Survey Data

The happiness index is a composite indicator used to measure the subjective well-being of individuals or populations. Unlike conventional economic indicators, the happiness index reflects an individual's overall perception of their quality of life, encompassing emotional, social, health, and economic aspects. Many countries and international organizations have developed happiness indices, such as the World Happiness Report, which relies on survey data from the Gallup World Poll [18].

Happiness data is generally collected through surveys that include various types of questions, such as Likert scale items, open-ended questions, and life satisfaction scores. A primary challenge in processing happiness survey data lies in the large number of subjective and highly correlated variables [19]. Therefore, analytical techniques are required to filter out information that is truly relevant for constructing a valid and reliable index. In this context, feature selection and feature reduction play a central role in accurately filtering and reducing data dimensionality.

C. Role of Feature Selection

Feature selection plays a crucial role in the development of efficient and accurate predictive models, particularly in social domains such as happiness measurement. By selecting only the most relevant features, the model not only becomes simpler in structure but also more robust in identifying meaningful patterns [20]. In the context of happiness indices, feature selection enables researchers to filter out the variables that best represent dimensions of happiness, such as social relationships, mental health, financial conditions, and trust in government [21].

Additionally, feature selection helps reduce bias and multicollinearity among variables, thereby improving the validity and reliability of the analysis results [17]. With the proper approach, this process can yield predictive models that are not only superior in quantitative performance but also substantively interpretable by policymakers or social researchers [22]. Therefore, the integration of feature selection techniques becomes a crucial component in the analysis of complex and multidimensional happiness survey data [23].

II. METHODOLOGY

This study employs a Systematic Literature Review (SLR) approach to examine the role and effectiveness of feature selection and feature reduction methods in the analysis of survey-based happiness indices. The review procedure follows the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020, designed to ensure transparency and replicability in the search and selection of scientific literature [24].

The application of the PRISMA methodology aims to enhance traceability and reproducibility of the review process, while minimizing potential selection bias by explicitly documenting each stage of the screening process [25]. By adhering to the PRISMA workflow, this study is expected to provide a more objective, comprehensive, and accountable literature synthesis, resulting in an accurate research mapping of feature selection and reduction techniques in the happiness domain [26]. This approach also supports the structured development of empirical evidence to answer the research questions posed.

A. Record Identification

This review analyzes 40 peer-reviewed studies from a curated collection focused on the application of machine learning in mental health. The studies were published between 2020 and 2025, ensuring contemporary relevance and covering the latest developments in the field. The curated collection was selected based on topic relevance, methodological quality, and geographic representativeness to provide a comprehensive overview of the current state of feature selection and feature reduction methods, particularly in the domain of happiness indices. Studies were identified through a systematic search across several major electronic databases, including PubMed, Scopus, IEEE Xplore, and Web of Science. Keywords used included combinations of terms such as "happiness index", "happiness indicator", "feature selection well-being", "dimensionality reduction survey", "feature selection dimensionality reduction", and other related terms. Of the total 295 articles identified in the initial search phase, 152 articles were selected after title and abstract evaluation. Full-text reviews were then conducted to assess suitability against inclusion criteria, ultimately resulting in 40 studies that met the eligibility requirements for inclusion in this review.

B. Record Selection

Article selection was carried out through a multi-stage process in accordance with the PRISMA 2020 guidelines, starting with an initial screening of titles and abstracts,

followed by a check for duplicates, and full-text evaluation based on eligibility criteria. This process aims to ensure that only relevant, credible, and high-quality literature is further analyzed in this study.

The inclusion criteria for this research include articles that have undergone peer review, are classified within Quartile rankings (Q1–Q4) based on journal rankings, were published between 2020 and 2025, are written in English, and are available in full-text format. Exclusion criteria consisted of articles that were duplicated across databases, came from journals outside the Quartile rankings (non-indexed/low quality), were not relevant to the main topic of feature selection or feature reduction, or could not be accessed or were not available in full-text format.

The application of these criteria resulted in a set of eligible and relevant articles to be further analyzed in order to systematically and responsibly answer the research questions. After completing the selection process, 40 studies met all inclusion criteria and were evaluated in this review to form the basis for the synthesis of findings. Figure 1. Figure caption.

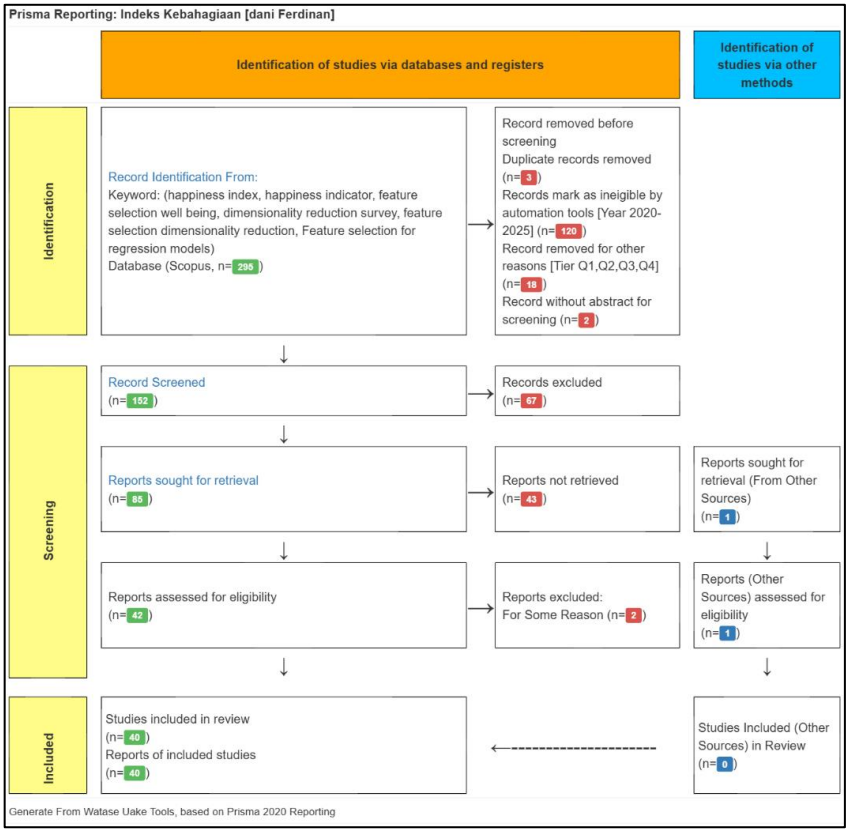


Figure 1. Prisma Report

C. Data Extraction

All articles obtained during the literature search were systematically evaluated to assess their relevance to the research problem and objectives. The evaluation was conducted through a thorough review of the full content of each article. Studies that did not meet the inclusion criteria or did not directly relate to the topic of predicting commodity prices were eliminated from further analysis. The initial selection process involved independent reviews of titles and abstracts by different reviewers to ensure objectivity and improve the validity of the literature selection.

Articles that met the final criteria were used as primary sources to identify various algorithmic approaches in the application of machine learning, examine the methodological challenges faced in the prediction process, and outline the direction and trends of future research in this field.

III. RESULT

A. General Statistics

To provide an initial overview of the characteristics of the literature analyzed in this review, a descriptive statistical analysis was performed on the 40 selected articles based on several key aspects, including the classification of research questions, publication year distribution, research design, methodological approach type, country distribution, and key terms.

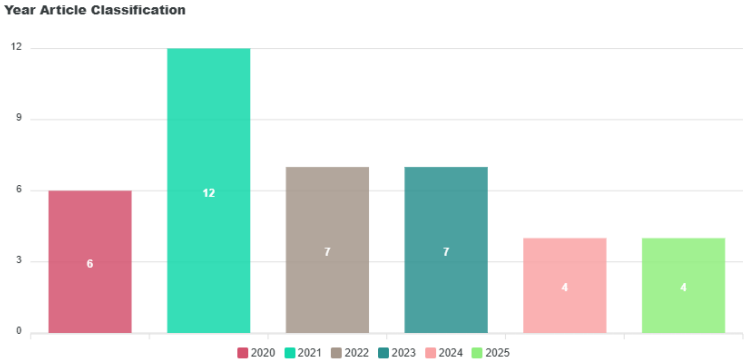


Figure 2. Years Distribution

Based on Figure 2, the publication year distribution shows an increasing trend in 2021, with a total of 12 articles, followed by 2022 and 2023 (with 7 articles each). The years 2020, 2024, and 2025 contributed between 4 and 6 articles each, reflecting a consistent interest in this topic over the past five years.

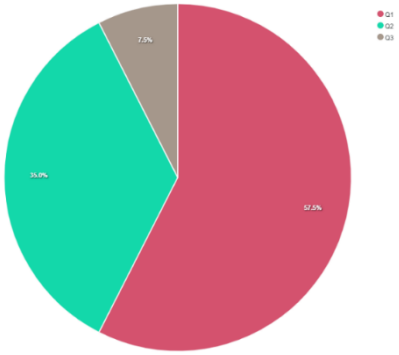


Figure 3. Journal Tier Distribution.

The distribution of the analyzed articles based on journal tier (Scopus Quartile), as shown in Figure 3, indicates that the majority are from Q1 journals (57.5%), reflecting the high quality and credibility of the publications reviewed. Additionally, 35% of the articles were published in Q2 journals, while only 7.5% were from Q3 journals.

This indicates that topics related to feature selection are frequently discussed in high-reputation journals, particularly those focusing on technical and methodological contributions based on machine learning. The dominance of Q1 journals further strengthens the position of this review as a relevant literature analysis grounded in reliable scientific sources.

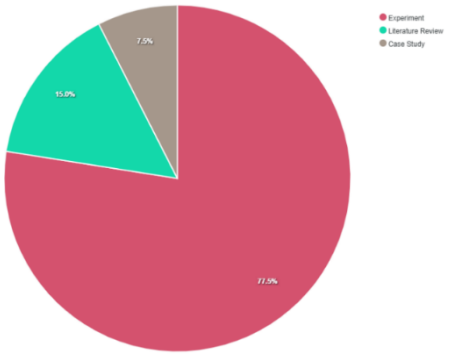


Figure 4. Distribution of Research Methodologies

The majority of studies employed an experimental approach (77.5%), reflecting a strong focus on evaluating the performance of algorithms and models based on survey data or simulations, as shown in Figure 4. Meanwhile, 15% of the articles were literature reviews, and 7.5% were case studies or context-based studies.

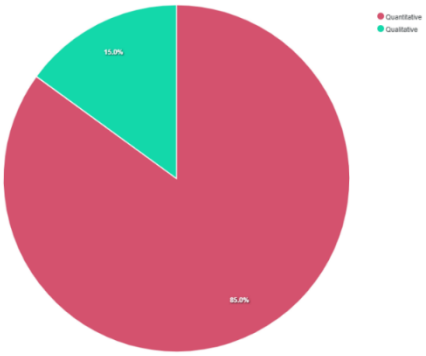


Figure 5. Research Design Distribution

The majority of articles fall into the quantitative category (85%), emphasizing empirical and computational testing, as shown in Figure 5. Qualitative approaches (15%) were used to explore questionnaire design, the interpretation of indicators, or non-numeric perceptions of happiness.

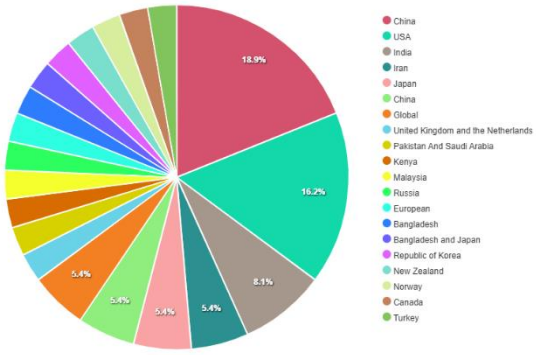
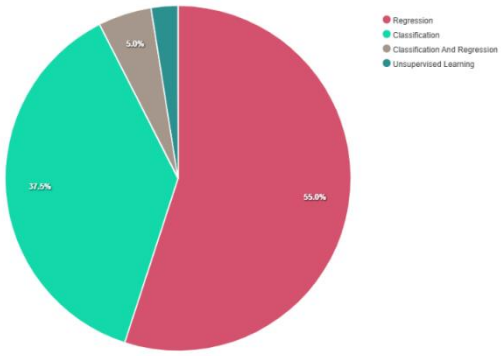


Figure 6. Geographic Distribution

The studies analyzed originate from various countries, with the highest representation coming from China (18.9%), followed by the USA (16.2%) and India (8.1%), as shown in Figure 6. Several articles also have a global or multinational scope, including studies from Europe, Pakistan, Korea, and other Asian countries

Figure 7. Meta Word Cloud.

The word cloud analysis, as shown in Figure 7, reveals that the most dominant keywords are "Feature selection", "Dimensionality reduction", "Machine learning", "Happiness", and "LASSO". This indicates that computational and statistical approaches are highly dominant in studies related to the measurement of happiness indices.



Figma 8. Distribution of Modeling Approaches.

As shown in Figure 8, the majority of studies in this analysis employ a regression approach (55%) to model happiness indices, primarily because the target variables in many cases are numeric, such as happiness scores or levels. Classification approaches are also widely used (37.5%), especially when happiness data is categorized into specific classes such as high, medium, or low. A small number of studies combine both approaches (5%) or apply unsupervised learning methods (2.5%) to identify hidden patterns in data without clear target variables. These findings suggest that the choice of modeling method is generally tailored to the characteristics of the data and the objectives of the analysis.

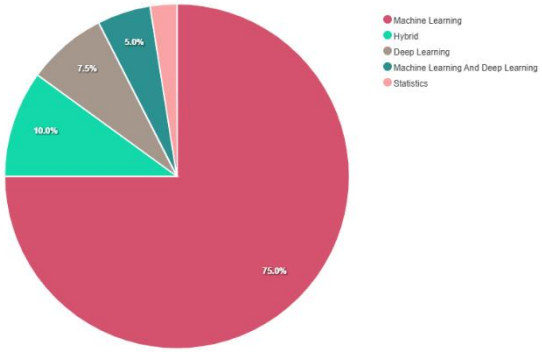


Figure 9. Distribution of Machine Learning Approaches

The majority of articles in this review implement a Machine Learning approach (75%) as the foundation for feature selection, such as Random Forest, XGBoost, and Support Vector Machine. Some studies also utilize Hybrid approaches (10%) and Deep Learning (7.5%), while a combination of Machine Learning and Deep Learning is found in 5% of the articles. Only a small number of articles use conventional statistical approaches (2.5%), such as stepwise regression without algorithmic modeling.

This indicates that machine learning approaches dominate the research landscape related to feature selection in happiness survey data, which aligns with the growing trend of using data-driven modeling in quantitative social sciences, as shown in Figure 9.

B. Findings Based on RQ

RQ 1: How do feature selection methods affect the accuracy of happiness index measurement using survey data?

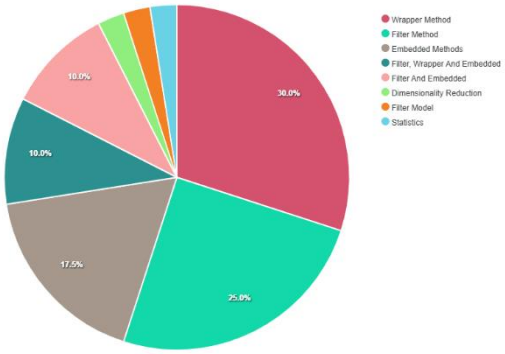


Figure 10. Distribution of Feature Selection Methods

Three main feature selection approaches are widely used in the studies analyzed, as illustrated in Figure 10. Wrapper methods occupy the largest share (30%) due to their ability to achieve high accuracy by directly evaluating feature subset combinations against model performance. On the other hand, filter methods remain popular (25%) because they are computationally lightweight and can be quickly applied without relying on a specific modeling algorithm. Meanwhile, embedded methods are used in about 17.5% of studies, particularly in models like Lasso Regression and decision tree algorithms, which perform feature selection internally during the training process. Additionally, some studies combine multiple approaches, such as filter-wrapper-embedded (10%), and more specific methods like dimensionality reduction (3%), filter models (2.5%), and basic statistics (2.5%). These findings indicate that, although wrapper approaches remain dominant, the variety of feature selection techniques continues to evolve according to the needs of the model and the characteristics of survey data.

Based on the understanding of the diverse feature selection approaches presented in Figure 5, it can be concluded that feature selection methods play a significant role in improving the prediction accuracy of happiness index measurement using survey data. Correlation-based feature selection methods, such as Correlation-based Feature Selection (CFS), have consistently delivered higher accuracy compared to Principal Component Analysis (PCA), especially in binary and multi-label classification using fully connected neural network (FCNN) models [27]. Additionally, ensemble methods like XGBoost outperform Random Forest and Lasso Regression in predicting happiness indices, with R^2 values reaching 85.03% [18].

The Recursive Feature Elimination (RFE) approach combined with Random Forest also shows superior performance ($R^2 = 0.72$) compared to other methods, such as Gray Relational Analysis (GRA) and Elastic Net [28]. Other studies emphasize the importance of explicit variable selection, which significantly improves the prediction accuracy of

happiness indices through models like Support Vector Machine (SVM) and Random Forest, with R^2 approaching 0.99 [29].

Furthermore, the Convex Least Angle Regression LASSO (CLAR-LASSO) method demonstrates a performance improvement in classification, achieving an accuracy of 95.67%, surpassing conventional methods such as Pearson correlation, LASSO, and LAR LASSO [30]. These studies indicate that optimal feature selection through integrated approaches can provide high prediction accuracy for survey or socio-economic data.

RQ 2: To what extent do feature selection methods improve the validity and reliability of variables in survey-based happiness measurement?

The use of feature selection methods significantly enhances the validity and reliability of variables in happiness measurement. A study by Jaiswal and Gupta demonstrates that Random Forest achieves an accuracy of 92.27%, highlighting its ability to significantly improve the validity of happiness indicators through the selection of relevant attributes [18]. The Enhanced Whale Optimization Algorithm (EWOA) also shows high accuracy, with a Mean Absolute Error (MAE) below 1% and a Root Mean Square Error (RMSE) below 1.3% in battery health prediction studies, affirming the relevance of optimization-based feature selection methods [31].

Another study using the ReliefF-based initialization approach shows more stable and accurate feature selection in genetic and health data, consistently improving the validity of selected variables [32]. Moreover, the combination of Recursive Feature Elimination with Random Forest (RFE-RF) also demonstrates a significant improvement in indicator validity by retaining relevant multidimensional variables [33].

Approaches like CLAR-LASSO enhance the validity of indicators such as "venture capital received", "funding rounds", and "angel investment", demonstrating the relevance of features in the socio-economic context [30]. Graph-based feature selection methods have also proven effective in generating features with strong associations and high validity in biological contexts, supporting their use in high-dimensional happiness index survey data [34].

RQ 3: What is the distribution of feature handling methods, and which is the most dominant?

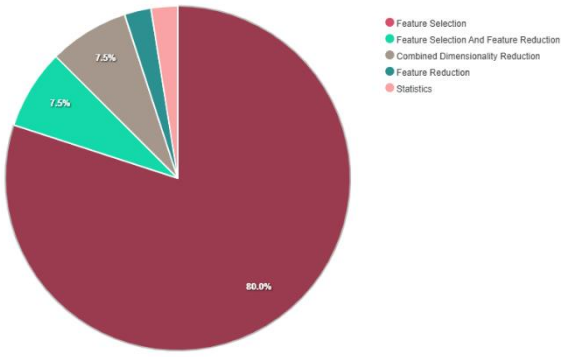


Figure 11. Distribution of Feature Processing Approaches.

As illustrated in Figure 11, 80% of the studies explicitly apply feature selection, emphasizing the primary focus of research on selecting important variables. Some studies also combine feature selection and feature reduction (7.5%), or use pure reduction methods (7.5%), such as PCA. Very few studies combine all approaches or only use basic statistics (2.5%), indicating that algorithm-based reduction is preferred in this domain.

This trend, as depicted in Figure 5, is further reinforced by specific findings that the combination of feature selection and feature reduction demonstrates high effectiveness in

improving the performance of happiness index analysis. For example, the combination of XGBoost and ElasticNet has proven effective in improving real-time prediction accuracy of happiness indices based on Google Trends™ data, yielding excellent metrics (MAE = 0.069, RMSE = 0.0938) [1].

Within the realm of feature selection, methods based on regularized regression are the most dominant. A studies leverage models such as LASSO, Ridge, or Elastic Net for this purpose. These techniques are frequently chosen for their innate ability to shrink the coefficients of irrelevant features to zero, thus performing an automatic and effective feature selection [35]. For instance, one study successfully used a Resampling-based Lasso Feature Selection (RLFS) to improve the accuracy of cancer prediction from high-dimensional gene data [36]. Another study identified 155 critical features from over a thousand using LASSO to predict RNA cleavage efficiency, achieving a Pearson correlation coefficient of 0.74 [37]. The proposed Feature Selection with Orthogonal Regression (FSOR) method was also shown to outperform the classic LASSO approach [38].

Beyond regression-based methods, other techniques are also utilized. Filter methods like the F-test, Neighborhood Component Analysis (NCA), and RReliefF were applied in a study on the design of concrete beams, where NCA proved superior for training the prediction model [39]. Furthermore, some studies propose novel methods, such as an adaptive graph-based model for unsupervised feature selection (AGUFS), which excels in clustering tasks by simultaneously selecting uncorrelated and discriminative features [40]. In hydrological forecasting, both automatic selection (Greedy Forward Selection) and manual selection based on domain expertise were explored, showing that the best approach can depend on the specific characteristics of the dataset [41]. The use of Hesitant Fuzzy Correlation to enhance regularized regression models also showed promise in improving feature selection quality by maximizing relevance and minimizing redundancy [42].

A synthesis of the reviewed literature is summarized in Table 1, which shows the relationship between research questions, methods used, and the effects of applying feature selection techniques on the accuracy and validity of happiness measurement

Table 1. Summary of Study Findings Based on Research Questions (RQ)

RQ	Research Question	Key Findings	Dominant Methods/Models	Main Effects
RQ1	How do feature selection methods affect the accuracy of happiness index measurement using survey data?	Feature selection methods such as wrapper, filter, and embedded methods have been proven to enhance model prediction accuracy.	XGBoost, Random Forest, Lasso, CLAR-LASSO, SVM, RFE	Increases R^2, F1-score, and accuracy; reduces prediction errors; prevents overfitting.
RQ2	To what extent do feature selection methods improve the validity and reliability of variables in survey-based happiness measurement?	FS strengthens the validity of indicators and the reliability between models, especially in maintaining the stability and semantic relationships of features.	Random Forest, CLAR-LASSO, ReliefF, Graph-based FS, EPPSFS	Filters relevant features theoretically; maintains feature stability across tests; avoids redundancy.
RQ3	What is the distribution of feature handling methods, and which is the most dominant?	feature selection is the most dominant method.	LASSO, Ridge, Elastic Net	Identify important variables

Table 1 presents a synthesis of the key findings from the literature studies analyzed based on the three main research questions (RQs). Each row summarizes how feature

selection (FS) and feature reduction (FR) methods impact happiness index measurement, including the dominant models/algorithms used and the main effects achieved. The findings indicate that FS approaches (wrapper, filter, embedded) and the combination of FS-FR contribute to improvements in accuracy, model efficiency, and interpretability, particularly in the context of survey data and high-dimensional data

IV. DISCUSSION

This discussion commences with the most crucial finding of our Systematic Literature Review (SLR): a conspicuous absence of any articles that explicitly integrate feature selection or reduction methods with the analysis of a happiness index. A systematic and extensive search across prominent academic databases, including Scopus, ScienceDirect, and IEEE Xplore, using a dedicated SLR platform, yielded zero results for this specific intersection. This lacuna in the literature confirms a significant research gap and serves as the primary justification for this study, underscoring its novelty and urgency. Consequently, to provide a foundational roadmap for future research, this review synthesizes findings from analogous fields where these techniques are intensively applied such as biology, medicine, and chemistry to extrapolate methodological relevance for the context of happiness measurement.

The use of various feature selection approaches (wrapper, filter, embedded, and hybrid) provides a deeper understanding of their impact on prediction accuracy and the validity of the variables used.

In RQ1, wrapper methods proved to be the most dominant due to their ability to achieve high accuracy, although they are computationally expensive. Methods such as Recursive Feature Elimination (RFE), XGBoost, and CLAR-LASSO showed excellent prediction performance, with high R^2 values and classification accuracy approaching 96%. This indicates that proper feature selection can reduce noise and improve the reliability of predictive models in the social domain.

For RQ2, the focus was on the validity and reliability of variables. Methods such as Random Forest, CLAR-LASSO, ReliefF, and EWOA showed that with the selection of relevant features, happiness indicators could be more valid both conceptually and empirically. For example, feature selection based on correlation and the strength of relationships between variables helps identify indicators that truly reflect happiness, rather than mere confounding variables.

Meanwhile, RQ3 highlights the distribution of feature handling methods is heavily skewed towards feature selection, with regularized regression emerging as the dominant technique, likely due to its inherent efficiency in automatically identifying and isolating the most impactful predictors from high-dimensional data.

This discussion indicates that the choice of feature selection methods cannot be made generically. Instead, the approach used must be tailored to the type of data, the analysis objectives, and the complexity of the variables being analyzed. The more complex and multidimensional the data, the greater the benefits of hybrid and integrated approaches in feature selection.

V. CONCLUSION

In conclusion, the primary contribution of this study is the definitive identification of a critical research gap: a complete absence of literature directly applying feature selection and reduction techniques to the analysis of happiness indices. To address this gap, this review systematically synthesized methodologies from analogous domains where these techniques are mature, such as biology, medicine, and chemistry. Our synthesis reveals that approaches like wrapper methods are dominant in these fields due to their high predictive accuracy, while techniques such as regularized regression are prevalent for their efficiency in handling high-dimensional data. These findings establish

methodological benchmarks that can guide the first wave of empirical work in happiness analytics, ultimately contributing to more reliable and efficient measurement frameworks.

The primary limitation of this work is inherent to the identified research gap; our conclusions are necessarily extrapolated from other fields and await empirical validation within the specific context of social and psychological data related to happiness. Furthermore, the reviewed literature from these other domains shows considerable variation in methods and a general disconnect between statistical optimization and psychometric theory, such as construct validity, which highlights challenges that will likely need to be addressed in future happiness research. Building upon these findings, the foremost priority for future research is to conduct foundational empirical studies that apply feature selection methods to diverse and representative happiness datasets. The dominant techniques identified in this review, such as wrapper methods and regularized regression, serve as promising starting points. Moreover, establishing methodological standards, particularly for evaluating construct validity, is crucial to bridge the gap between computational optimization and psychological measurement theory. The development of hybrid feature selection frameworks that combine machine learning efficiency with domain-specific interpretability could advance both the technical performance and practical utility of happiness index modeling. Such efforts will not only deepen the understanding of relevant predictors but also enhance the robustness and policy relevance of happiness measurement in diverse sociocultural contexts.

REFERENCES

[1] T. Greyling and S. Rossouw, "Development and Validation of a Real-Time Happiness Index Using Google Trends™," *J Happiness Stud*, vol. 26, no. 3, Mar. 2025, doi: 10.1007/s10902-025-00881-9.

[2] J. Vendrow, J. Haddock, D. Needell, and L. Johnson, "Feature selection from lyme disease patient survey using machine learning," *Algorithms*, vol. 13, no. 12, Dec. 2020, doi: 10.3390/a13120334.

[3] M. Jamei, A. S. Mohammed, I. Ahmadianfar, M. M. S. Sabri, M. Karbasi, and M. Hasanipanah, "Predicting Rock Brittleness Using a Robust Evolutionary Programming Paradigm and Regression-Based Feature Selection Model," *Applied Sciences (Switzerland)*, vol. 12, no. 14, Jul. 2022, doi: 10.3390/app12147101.

[4] P. Agrawal, H. F. Abutarboush, T. Ganesh, and A. W. Mohamed, "Metaheuristic algorithms on feature selection: A survey of one decade of research (2009-2019)," *IEEE Access*, vol. 9, pp. 26766–26791, 2021, doi: 10.1109/ACCESS.2021.3056407.

[5] O. Manullang, C. Prianto, and N. H. Harani, "Analisis Sentimen Untuk Memprediksi Hasil Calon Pemilu Presiden Menggunakan Lexicon Based dan Random Forest," *Jurnal Ilmiah Informatika (JIF)*, vol. 11, no. 02, pp. 160–169, 2023, doi: https://doi.org/10.33884/jif.v11i02.7987.

[6] R. Jain and W. Xu, "RHDSI: A novel dimensionality reduction based algorithm on high dimensional feature selection with interactions," *Inf Sci (N Y)*, vol. 574, pp. 590–605, Oct. 2021, doi: 10.1016/j.ins.2021.06.096.

[7] H. Gunduz, "An efficient dimensionality reduction method using filter-based feature selection and variational autoencoders on Parkinson's disease classification," *Biomed Signal Process Control*, vol. 66, Apr. 2021, doi: 10.1016/j.bspc.2021.102452.

[8] A. A. Aouragh, M. Bahaj, and F. Toufik, "Diabetes Prediction: Optimization of Machine Learning through Feature Selection and Dimensionality Reduction," *International journal of online and biomedical engineering*, vol. 20, no. 8, pp. 100–114, May 2024, doi: 10.3991/ijoe.v20i08.47765.

[9] H. B. Andrews, L. R. Sadergaski, and S. K. Cary, "Pursuit of the Ultimate Regression Model for Samarium(III), Europium(III), and LiCl Using Laser-

Induced Fluorescence, Design of Experiments, and a Genetic Algorithm for Feature Selection," *ACS Omega*, vol. 8, no. 2, pp. 2281–2290, Jan. 2023, doi: 10.1021/acsomega.2c06610.

[10] T. Wang, "A combined model for short-term wind speed forecasting based on empirical mode decomposition, feature selection, support vector regression and crossvalidated lasso," *PeerJ Comput Sci*, vol. 7, pp. 1–23, 2021, doi: 10.7717/peerj-cs.732.

[11] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm Evol Comput*, vol. 54, May 2020, doi: 10.1016/j.swevo.2020.100663.

[12] D. Bender, D. J. Licht, and C. Nataraj, "A novel embedded feature selection and dimensionality reduction method for an SVM type classifier to predict periventricular leukomalacia (PVL) in neonates," *Applied Sciences (Switzerland)*, vol. 11, no. 23, Dec. 2021, doi: 10.3390/app112311156.

[13] P. Dhal and C. Azad, "A comprehensive survey on feature selection in the various fields of machine learning," *Applied Intelligence*, vol. 52, no. 4, pp. 4543–4581, Mar. 2022, doi: 10.1007/s10489-021-02550-9.

[14] D. A. Otchere, T. O. A. Ganat, J. O. Ojero, B. N. Tackie-Otoo, and M. Y. Taki, "Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions," *J Pet Sci Eng*, vol. 208, Jan. 2022, doi: 10.1016/j.petrol.2021.109244.

[15] M. R. Islam, A. A. Lima, S. C. Das, M. F. Mridha, A. R. Prodeep, and Y. Watanobe, "A Comprehensive Survey on the Process, Methods, Evaluation, and Challenges of Feature Selection," *IEEE Access*, vol. 10, pp. 99595–99632, 2022, doi: 10.1109/ACCESS.2022.3205618.

[16] M. Ashraf *et al.*, "A Survey on Dimensionality Reduction Techniques for Time-Series Data," *IEEE Access*, vol. 11, pp. 42909–42923, 2023, doi: 10.1109/ACCESS.2023.3269693.

[17] Y. Liu, P. Pi, and S. Luo, "A semi-parametric approach to feature selection in high-dimensional linear regression models," *Comput Stat*, vol. 38, no. 2, pp. 979–1000, Jun. 2023, doi: 10.1007/s00180-022-01254-z.

[18] S. Çelik, B. Doğanlı, M. Ü. Şaşmaz, and U. Akkucuk, "Accuracy Comparison of Machine Learning Algorithms on World Happiness Index Data," *Mathematics*, vol. 13, no. 7, Apr. 2025, doi: 10.3390/math13071176.

[19] A. Stelmokienė and G. Jarašiūnaitė-Fedosejeva, "Is Leadership Position Related To More Social Inclusion, Happiness, And Satisfaction With Life? The Importance Of Power Distance Index," *Business: Theory and Practice*, vol. 24, no. 1, pp. 148–159, Jan. 2023, doi: 10.3846/btp.2023.16705.

[20] M. Li, H. Wang, L. Yang, Y. Liang, Z. Shang, and H. Wan, "Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction," *Expert Syst Appl*, vol. 150, Jul. 2020, doi: 10.1016/j.eswa.2020.113277.

[21] C. H. Feng, M. L. Disis, C. Cheng, and L. Zhang, "Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial logistic regression models," *Laboratory Investigation*, vol. 102, no. 3, pp. 236–244, Mar. 2022, doi: 10.1038/s41374-021-00662-x.

[22] S. Li, J. Yu, H. Kang, and J. Liu, "Genomic Selection in Chinese Holsteins Using Regularized Regression Models for Feature Selection of Whole Genome Sequencing Data," *Animals*, vol. 12, no. 18, Sep. 2022, doi: 10.3390/ani12182419.

[23] J. J. A. Mendes Junior, M. L. B. Freitas, H. V. Siqueira, A. E. Lazzaretti, S. F. Pichorim, and S. L. Stevan, "Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels

armband approach," *Biomed Signal Process Control*, vol. 59, May 2020, doi: 10.1016/j.bspc.2020.101920.

[24] Y. Lyu, Y. Feng, and K. Sakurai, "A Survey on Feature Selection Techniques Based on Filtering Methods for Cyber Attack Detection," Mar. 01, 2023, *MDPI*. doi: 10.3390/info14030191.

[25] B. Zuhri and N. H. Harani, "Studi Literatur: Optimasi Algoritma Machine Learning Untuk Prediksi Penerimaan Mahasiswa Pascasarjana," *Jurnal Informatika dan Teknologi Komputer (J-ICOM)*, vol. 05, no. 01, pp. 01–10, 2024, doi: https://doi.org/10.55377/j-icom.v5i1.8074.

[26] X. Chen, X. Zhu, Y. Lu, and Z. Pu, "Non-negative low-rank adaptive preserving sparse matrix regression model for supervised image feature selection and classification," *IET Image Process*, vol. 17, no. 7, pp. 2056–2071, May 2023, doi: 10.1049/ipr2.12772.

[27] F. Hassan *et al.*, "A hybrid approach for intrusion detection in vehicular networks using feature selection and dimensionality reduction with optimized deep learning," *PLoS One*, vol. 20, no. 2 February, Feb. 2025, doi: 10.1371/journal.pone.0312752.

[28] S. Zhang, J. Zhao, J. Yang, J. Xie, and Z. Sun, "Feature Selection and Regression Models for Multisource Data-Based Soil Salinity Prediction: A Case Study of Minqin Oasis in Arid China," *Land (Basel)*, vol. 13, no. 6, Jun. 2024, doi: 10.3390/land13060877.

[29] K. Kaushik, A. Bhardwaj, A. Aggarwal, and M. Kumar, "Enumerating happiness index during COVID-19 lockdowns using artificial intelligence techniques," *International Journal of Technology Management and Sustainable Development*, vol. 22, pp. 35–52, May 2023, doi: 10.1386/tmsd_00066_1.

[30] R. Allu and V. N. R. Padmanabhuni, "Convex Least Angle Regression Based LASSO Feature Selection and Swish Activation Function Model for Startup Survival Rate," *Cybernetics and Information Technologies*, vol. 23, no. 4, pp. 110–127, Nov. 2023, doi: 10.2478/cait-2023-0039.

[31] R. Wang *et al.*, "State of Health Estimation for Lithium-Ion Batteries Using Enhanced Whale Optimization Algorithm for Feature Selection and Support Vector Regression Model," *Processes*, vol. 13, no. 1, Jan. 2025, doi: 10.3390/pr13010158.

[32] B. Ahadzadeh *et al.*, "Improved binary differential evolution with dimensionality reduction mechanism and binary stochastic search for feature selection," *Appl Soft Comput*, vol. 151, Jan. 2024, doi: 10.1016/j.asoc.2023.111141.

[33] R. Zhu *et al.*, "Well-Production Forecasting Using Machine Learning with Feature Selection and Automatic Hyperparameter Optimization," *Energies (Basel)*, vol. 18, no. 1, Jan. 2025, doi: 10.3390/en18010099.

[34] C. Gakii, P. O. Mireji, and R. Rimiru, "Graph Based Feature Selection for Reduction of Dimensionality in Next-Generation RNA Sequencing Datasets," *Algorithms*, vol. 15, no. 1, Jan. 2022, doi: 10.3390/a15010021.

[35] M. Mokhtia, M. Eftekhari, and F. Saberi-Movahed, "Dual-manifold regularized regression models for feature selection based on hesitant fuzzy correlation," *Knowl Based Syst*, vol. 229, Oct. 2021, doi: 10.1016/j.knosys.2021.107308.

[36] A. R. Patil and S. Kim, "Combination of ensembles of regularized regression models with resampling-based lasso feature selection in high dimensional data," *Mathematics*, vol. 8, no. 1, Jan. 2020, doi: 10.3390/math8010110.

[37] D. Ueno, H. Kawabe, S. Yamasaki, T. Demura, and K. Kato, "Feature selection for RNA cleavage efficiency at specific sites using the LASSO regression model in Arabidopsis thaliana," *BMC Bioinformatics*, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s12859-021-04291-5.

[38] B. Tang, Y. Wang, Y. Chen, M. Li, and Y. Tao, "A Novel Early-Stage Lung Adenocarcinoma Prognostic Model Based on Feature Selection With Orthogonal Regression," *Front Cell Dev Biol*, vol. 8, Jan. 2021, doi: 10.3389/fcell.2020.620746.

[39] W. K. Hong and T. D. Pham, "Reverse designs of doubly reinforced concrete beams using Gaussian process regression models enhanced by sequence training/designing technique based on feature selection algorithms," *Journal of Asian Architecture and Building Engineering*, vol. 21, no. 6, pp. 2345–2370, 2022, doi: 10.1080/13467581.2021.1971999.

[40] Y. Huang, Z. Shen, F. Cai, T. Li, and F. Lv, "Adaptive graph-based generalized regression model for unsupervised feature selection," *Knowl Based Syst*, vol. 227, Sep. 2021, doi: 10.1016/j.knosys.2021.107156.

[41] V. Moreido, B. Gartsman, D. P. Solomatine, and Z. Suchilina, "How well can machine learning models perform without hydrologists? Application of rational feature selection to improve hydrological forecasting," *Water (Switzerland)*, vol. 13, no. 12, Jun. 2021, doi: 10.3390/w13121696.

[42] M. Mokhtia, M. Eftekhari, and F. Saberi-Movahed, "Feature selection based on regularization of sparsity based regression models by hesitant fuzzy correlation," *Applied Soft Computing Journal*, vol. 91, Jun. 2020, doi: 10.1016/j.asoc.2020.106255.