Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 14 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i14.56383

PAPER

Depression Severity Prediction Using Stacked Heterogenous Learning on Integrated Biomedical and Lifestyle Data

Adefemi Ayodele[1], Adedotun Adetunla[2] (✉), Esther Akinlabi[3]

[1]University of East London, London, United Kingdom

[2]University of Johannesburg, Johannesburg, South Africa

[3]Colorado State University, Fort Collins, CO, USA

dotunadetunla@gmail.com

ABSTRACT

Depression, a major global health concern, requires improved methods for timely and precise assessment. This study proposes a stacked heterogeneous ensemble framework to predict depression severity, measured by cumulative PHQ-9 scores, using multimodal data from the 2013–2014 NHANES survey. The dataset integrates six domains, such as demographics, diet, physical examination, laboratory results, medication use, and mental health responses, capturing both biomedical and lifestyle indicators. Seven base learners (SVR, Ridge, ElasticNet, KNN, Decision Tree, LightGBM, and XGBoost) were stacked in two stages. First-level outputs were fused with a Voting Regressor (Gradient Boosting Regressor + Linear Regression). The PHQ-9 scale ranges from 0 to 27, with higher values reflecting greater severity. A total of 19,560 participants were analyzed (80/20 split: 15,648 train, 3,912 test), with inverse-frequency weighting to address class imbalance. Results show the Stacking Regressor outperformed single models, achieving MSE = 1.66, RMSE = 1.29, and R^2 = 0.93. These findings highlight ensemble learning's promise for psychiatric modeling and its potential as a scalable, interpretable tool for clinical decision support.

KEYWORDS

biomedical data, depression severity, mental health informatics, PHQ-9, stacked ensemble learning

1 INTRODUCTION

Depression ranks among the world's most debilitating conditions, affecting over 300 million people globally and costing the United States alone more than $210 billion each year in lost productivity and healthcare expenditures [1]. Early identification and continuous monitoring of depressive symptom severity are essential for prompt intervention and better patient outcomes. However, traditional tools such as the Patient Health Questionnaire-9 (PHQ-9) and the clinician-rated Hamilton Depression

Ayodele, A., Adetunla, A., Akinlabi, E. (2025). Depression Severity Prediction Using Stacked Heterogenous Learning on Integrated Biomedical and Lifestyle Data. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(14), pp. 156–167. https://doi.org/10.3991/ijoe.v21i14.56383

Article submitted 2025-05-13. Revision uploaded 2025-09-06. Final acceptance 2025-09-22.

Rating Scale (HAMD-17) are limited by subjective interpretation, recall inaccuracies, and inconsistent administration schedules [2]. These challenges have intensified since the COVID-19 pandemic, driving up the incidence of depression and the need for scalable, real-time evaluation methods [3]. Recent advances in machine learning (ML) holds promise for overcoming these challenges by leveraging large, multimodal datasets to generate objective risk scores. However, extant ML approaches to depression prediction have encountered three principal limitations, which are reliance on single-algorithm architectures that fail to exploit complementary strengths across model classes; another being narrow data scopes that omit either biomedical biomarkers or dynamic behavioral traces; and lastly, inadequate handling of class imbalance, particularly for severe cases [4]. For example, the PSYCHE-D model used wearable-device accelerometry and LightGBM to predict three-month changes in PHQ-9 with only 55.4% sensitivity, in part due to limited algorithmic diversity and no explicit imbalance mitigation [5]. Primary-care risk models incorporating demographic and chronic-illness covariates achieved moderate discrimination (c-statistic = 0.74) but relied entirely on static self-reports [6]. Smartphone-based pipelines improved F1 scores by 0.09 through passive sensor features, yet they neglected critical laboratory biomarkers [7]. Other efforts to address imbalance, such as combining feature-group partitioning with SMOTE, achieved up to 92.8% balanced accuracy in student samples, but without objective biomarker validation or broader population representativeness [8]. Even NHANES-derived classifiers boasting AUCs near 0.99 have largely focused on cross-sectional case–control discrimination rather than longitudinal forecasting [9]. Some other recent studies have predicted depression severity using various machine learning algorithms [10], [11], [12]. Collectively, these studies reveal three critical shortcomings, which are overdependence on homogeneous ML architectures, underutilization of biomarker-behavioral synergies, and lastly, insufficient methodological rigor in hyperparameter optimization and generalizability.

To address these limitations, this study introduces an ML framework grounded in heterogeneous ensemble learning, multimodal data integration, and systematic hyperparameter tuning, a stacked heterogeneous learning on integrated biomedical and lifestyle data, for predicting depression severity as measured by cumulative PHQ-9 scores. Drawing on the publicly available 2013–2014 NHANES dataset (Kaggle), the study integrates six complementary modules; demographic, dietary, examination, laboratory, medication, and questionnaire, into a unified analytical platform. This dataset captures both static biomedical markers (e.g., CRP, blood lead) and lifestyle indicators (e.g., sleep duration, physical activity), enabling a holistic view of biopsychosocial determinants.

The proposed framework brings three key innovations into a single, seamless pipeline. First, the study constructs a heterogeneous ensemble by combining seven distinct learning paradigms, including ElasticNet, SVR, KNN, Decision Tree, LightGBM, XGBoost, and Ridge Regression, within a two-stage stacking regressor whose meta-learner is a voting regressor (gradient boosting plus linear regression). Second, the study fuses NHANES's richly multimodal feature space covering demographic, dietary, examination, laboratory, medication, and questionnaire modules by performing mean imputation for missing values, encoding categorical variables, applying a combined F-test and mutual-information selector to retain the top 100 predictors, and scaling all features to the [0,1] interval. Third, to address the inherently skewed PHQ-9 score distribution, the study implemented stratified train/test splitting alongside inverse-frequency sample weighting during model fitting, ensuring that observations with severe depressive symptoms exert proportionally greater influence on the loss function and thereby improving our sensitivity to high-severity cases. The approaches adopted in this study equip clinicians with a dynamic tool to stratify risk, monitor progression, and tailor

interventions addressing a critical need in post-pandemic mental health care. The methodology and results will be discussed extensively.

2 METHODOLOGY

This section outlines the methodological framework, including the data sources and integration, preprocessing steps, feature engineering and selection strategies, model development, and evaluation metrics employed in the study. Figure 1 shows the architecture of the developed model.

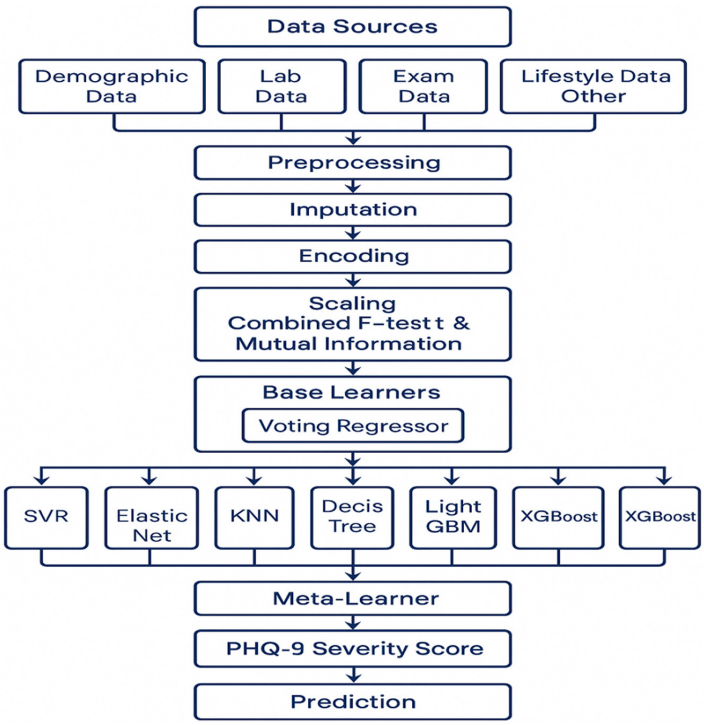

Fig. 1. Flowchart of the model

2.1 Data acquisition and integration

The 2013–2014 National Health and Nutrition Examination Survey (NHANES), a nationally representative dataset managed by the U.S. Centers for Disease Control and Prevention, was obtained from Kaggle. NHANES integrates multi-modal health data through interviews, physical examinations, and laboratory tests, offering a robust foundation for investigating depression's biopsychosocial determinants [13]. Six data modules were selected for their theoretical relevance to mental health outcomes:

- Demographics (demographic.csv): age, sex, household income, education, ethnicity.
- Dietary intake (diet.csv): macro- and micronutrient consumption patterns.
- Physical examination (examination.csv): anthropometrics (e.g., BMI) and vital signs (e.g., blood pressure).
- Laboratory results (labs.csv): clinical biomarkers such as fasting glucose and lipid profiles.
- Medication history (medications.csv): records of prescribed drugs.

- Self-reported questionnaire (questionnaire.csv): PHQ-9 items and other mental-health indicators.

Data integration was performed in Python using pandas, with files merged on the unique participant identifier SEQN to create a unified dataset. Non-standard file encodings were resolved using the chardet library. The merged dataset captured multidimensional predictors of depression, from biological markers to lifestyle factors. Figure 2 shows the schematic diagram of the data merging.

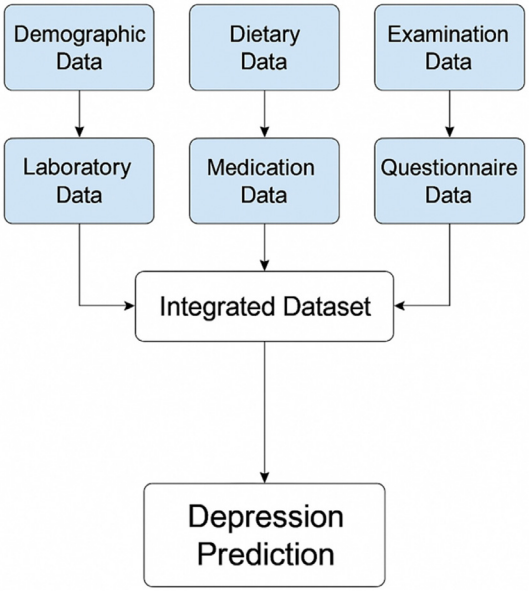

Fig. 2. Schematic diagram of the data aggregation

2.2 Data preprocessing and feature engineering

To ensure data quality and analytical validity, the dataset underwent a structured preprocessing and feature engineering pipeline. Initial data cleaning involved the removal of duplicate entries by verifying the uniqueness of SEQN identifiers. Missing numerical values were imputed using mean substitution, a method that retains sample size while preserving the central tendency. Outliers, such as records with implausible values like negative age, were filtered out to enhance the reliability of subsequent analysis. The target variable, depression severity (dpq_total), was constructed by summing individual PHQ-9 responses. Consistent with established PHQ-9 scoring guidelines [14], invalid codes (7 and 9) were treated as missing. Categorical variables—both nominal and ordinal—were transformed using one-hot encoding via pandas.get_dummies, and any residual NaN columns resulting from the encoding process were removed to prevent model training errors. Given the dataset's high dimensionality, a hybrid feature selection strategy was employed to optimize model performance. A custom CombinedFeatureSelector was developed to incorporate univariate F-test scores (via f_regression) alongside mutual information measures. These scores were normalized and combined equally to create a unified importance ranking [15]. The top 100 features most predictive of the dpq_total outcome were retained, effectively reducing dimensionality while preserving explanatory power (Saeys et al., 2007). Finally, all features were normalized using

MinMaxScaler, constraining values to the [0, 1] range to address scale disparities among variables such as BMI and blood glucose.

2.3 Model development and hyperparameter optimization

A suite of supervised regressors was trained to predict the dpq_total, covering linear, kernel-based, instance-based, and tree-based algorithmic paradigms.

Hyperparameters for each algorithm were optimized via 5-fold cross-validation using GridSearchCV, targeting minimization of negative mean squared error (MSE). This exhaustive search fine-tuned individual learners and capitalized on complementary model strengths; linear models provide baseline interpretability, while tree-based models capture complex interactions and set the stage for robust ensembling [16]. Table 1 provides the search spaces and rationales for each model.

Table 1. Hyperparameter search spaces and rationales for considered model

Model Type	Algorithm	Rationale	Hyperparameter Space
Linear	ElasticNet	Balance between L_1 and L_2 shrinkage	α ☐ {0.1, 1.0, 10.0}, l1_ratio ☐ {0.2, 0.5, 0.8}
Linear	Ridge Regression	Pure L_2 regularization	α ☐ {0.1, 1.0, 10.0}, solver ☐ {auto, sag, lsqr}
Kernel-Based	Support Vector Regression (SVR)	Capture non-linear relationships	C ☐ {0.1, 1.0, 10.0}, ε ☐ {0.01, 0.1, 0.5}, kernel ☐ {linear, rbf}
Instance-Based	K-Nearest Neighbors (KNN)	Local neighborhood modeling	n_neighbors ☐ {3, 5, 10}, p ☐ {1, 2} (L_1 vs. L_2), weights ☐ {uniform, distance}
Tree-Based	Decision Tree Regressor	Hierarchical, piecewise fitting	max_depth ☐ {3, 5, 7}, min_samples_split ☐ {2, 5, 10}
Boosting	LightGBM	Leaf-wise gradient boosting	n_estimators ☐ {100, 200, 300}, learning_rate ☐ {0.01, 0.1, 0.2}, num_leaves ☐ {31, 63, 127}
Boosting	XGBoost	Level-wise gradient boosting	n_estimators ☐ {100, 200, 300}, learning_rate ☐ {0.01, 0.1, 0.2}, max_depth ☐ {3, 5, 7}
Ensemble	Stacking Meta-Learner	Hierarchical recombination of predictions	Hybrid gradient boosting + linear combination

2.4 Ensemble learning and model evaluation

To enhance predictive robustness and generalization, two ensemble strategies were employed. The first was a simple voting regressor, combining the outputs of gradient boosting and linear regression to strike a balance between interpretability and flexibility. The second, a stacked ensemble, integrated first-level predictions from seven individually optimized models—SVR, Ridge, ElasticNet, KNN, decision tree, LightGBM, and XGBoost—into a second-level voting regressor. This hierarchical structure enabled the system to harness diverse model architectures, promoting interaction among different inductive biases for more stable and accurate predictions [17]. Model evaluation was conducted using an 80/20 stratified train-test split to maintain the distribution of depression severity levels. To mitigate class imbalance, particularly for underrepresented severe cases, inverse-frequency sample weighting was applied during training. Model performance was assessed using standard regression metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and Explained Variance Score (EVS). Comparative performance visualizations were generated using matplotlib and plotly, offering insights into trade-offs between predictive error and

explanatory power across different algorithms. All experiments were conducted in Google Colaboratory, utilizing Python 3 on a Google Compute Engine backend. The computational environment, equipped with 12.7 GB RAM and over 100 GB of storage, was sufficient for high-dimensional data ingestion, model training, and feature selection. Version control and environment snapshots ensured reproducibility of results throughout the development pipeline.

2.5 Ethical compliance and data access

NHANES protocols are reviewed by the NCHS Ethics Review Board. Public-use files are de-identified, with informed consent obtained at interview and exam. This study used only 2013–2014 de-identified data; no additional IRB approval was required. Data and documentation are available from CDC/NCHS.

3 RESULTS AND DISCUSSION

3.1 Model optimization and performance evaluation

The results demonstrate the effectiveness of ensemble learning techniques in predicting depression severity scores (PHQ-9) using a high-dimensional, engineered dataset. Individually, each base learner contributed unique strengths to the modeling process, informed by its inductive bias and capacity to capture specific data patterns. Among the single models, LightGBM and XGBoost stood out with superior performance, achieving high accuracy with low prediction error. LightGBM attained the best overall performance with a MSE of 1.78, a RMSE of 1.33, and a coefficient of determination (R^2) of 0.93, indicating strong generalization and robustness in mapping complex non-linear relationships.

Interestingly, traditional linear models such as ridge regression and ElasticNet underperformed in comparison, particularly ElasticNet, which posted the highest MSE (18.18) and lowest R^2 (0.25). This suggests that while regularized linear approaches offer interpretability and resilience against multicollinearity, they may be insufficient for capturing the intricate feature interactions present in this dataset. Similarly, the decision tree regressor, although more flexible, showed limited depth in pattern recognition with a modest R^2 of 0.62, highlighting the need for ensemble enhancement. The stacked regressor, integrating predictions from all seven base models into a second-level voting regressor, matched LightGBM's performance (MSE = 1.66, R^2 = 0.93), but with slightly better overall error distribution (RMSE = 1.29, MAE = 0.46). This validates the hypothesis that stacking multiple models with complementary strengths yields a more robust and stable prediction. The consistent performance gain achieved through stacking reinforces its utility in high-dimensional health data modeling, particularly when feature interactions are complex and non-linear. Moreover, SVR performed well compared to other traditional learners, posting an R^2 of 0.53. Despite its slightly higher error margin than tree-based models, SVR's ability to handle margin-based errors and non-linearities justifies its inclusion in the ensemble strategy. The KNN model also showed strong local pattern recognition (R^2 = 0.91), demonstrating its utility in approximating values where spatial similarity between data points is informative. The snapshot of the developed Python code is shown in Figure 3. Overall, the ensemble approach, particularly stacking, not only improved prediction accuracy but also mitigated the limitations of individual learners. The stratified train-test split and inverse-frequency sample weighting ensured fair representation of severe

depression cases, which are often underrepresented in mental health datasets. The model's performance metrics on the held-out test set are summarized in Table 2.

```
# Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as px
import chardet
import torch
import tensorflow as tf

from sklearn.model_selection import train_test_split, GridSearchCV, KFold
from sklearn.preprocessing import MinMaxScaler
from sklearn.feature_selection import SelectKBest, f_regression
from sklearn.ensemble import RandomForestRegressor, StackingRegressor,
GradientBoostingRegressor
from sklearn.linear_model import LinearRegression, Lasso
from sklearn.svm import SVR
from sklearn.neighbors import KNeighborsRegressor
from sklearn.metrics import r2_score, mean_squared_error,
explained_variance_score, mean_absolute_error
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from google.colab import drive
from sklearn.svm import SVR
from sklearn.linear_model import Ridge, ElasticNet, LinearRegression
from sklearn.neighbors import KNeighborsRegressor
from sklearn.tree import DecisionTreeRegressor
from xgboost import XGBRegressor
from lightgbm import LGBMRegressor
from sklearn.ensemble import (
    GradientBoostingRegressor,
    StackingRegressor,
    VotingRegressor
)
from sklearn.model_selection import GridSearchCV

# Mount Google Drive
```

```
# Add these to your existing model definitions

# 1. K-Nearest Neighbors (Instance-Based)
knn_param_grid = {
    'n_neighbors': [3, 5, 10],
    'weights': ['uniform', 'distance'],
    'p': [1, 2]  # L1 & L2 distances
}
knn_model = tune_model(KNeighborsRegressor(), knn_param_grid, X_train,
y_train)

# 2. ElasticNet (Blended Regularization)
en_param_grid = {
    'alpha': [0.1, 1.0, 10.0],
    'l1_ratio': [0.2, 0.5, 0.8]  # Mix of L1/L2
}
en_model = tune_model(ElasticNet(random_state=42), en_param_grid, X_train,
y_train)

# 3. Decision Tree (Non-Parametric)
dt_param_grid = {
    'max_depth': [3, 5, 7],
    'min_samples_split': [2, 5, 10]
}
dt_model = tune_model(DecisionTreeRegressor(random_state=42),
dt_param_grid, X_train, y_train)

# 4. Support Vector Regression (non-linear)
svr_param_grid = {
    'C': [0.1, 1, 10],
    'epsilon': [0.01, 0.1, 0.5],
    'kernel': ['linear', 'rbf']
}
svr_model = tune_model(SVR(), svr_param_grid, X_train, y_train)

# 5. Ridge Regression (linear)
ridge_param_grid = {
    'alpha': [0.1, 1.0, 10.0],
    'solver': ['auto', 'sag', 'lsqr']
```

Fig. 3. Snapshot of the stacked heterogeneous learning pipeline

Table 2. Performance evaluation results of the models

Model	MSE	RMSE	MAE	R^2	EVS
SVR	11.28	3.36	1.93	0.53	0.55
Ridge Regression	13.89	3.73	2.54	0.43	0.43
LightGBM	1.78	1.33	0.46	0.93	0.93
XGBoost	1.88	1.37	0.60	0.92	0.92
KNN	2.11	1.45	0.42	0.91	0.91
ElasticNet	18.18	4.26	2.96	0.25	0.25
Decision Tree	9.23	3.04	1.86	0.62	0.62
Stacking Regressor	1.66	1.29	0.46	0.93	0.93

3.2 Comparative analysis of model performance

The comparison of various machine learning algorithms for predicting depression severity, as measured by PHQ-9 scores, reveals distinct performance patterns that reflect each model's strengths and limitations when applied to complex, multimodal

mental health data. Ensemble and gradient boosting methods emerged as the most effective, showcasing their ability to handle the heterogeneity and high dimensionality inherent in psychological health datasets. Notably, both XGBoost and LightGBM delivered exceptional results, aligning with existing literature that underscores their strength in modeling non-linear relationships and managing diverse feature types. These models effectively captured intricate associations among behavioral, biological, and demographic variables. The KNN model also performed well, particularly due to its capacity to identify local similarities between cases. While not as flexible or scalable as boosting models, its instance-based learning mechanism demonstrated meaningful predictive power, especially when integrated into ensemble strategies where its limitations could be offset by more expressive models. In contrast, linear models such as Ridge Regression and ElasticNet underperformed, reflecting their limited ability to capture the interactive and non-linear patterns prevalent in mental health datasets. Their relatively low R^2 and high error rates suggest that while these models offer transparency, they may not be suited for complex prediction tasks involving diverse clinical and lifestyle indicators.

Support vector regression and decision tree regressors showed moderate results, performing better than linear models but still falling short of the ensemble approaches. Their interpretability and margin-based learning may offer utility in specific use cases but lacked the overall predictive strength required for robust clinical implementation in this context. Ultimately, the stacked ensemble approach delivered the best performance, leveraging the complementary strengths of all constituent models. This approach reinforced the value of hybrid modeling strategies in mental health prediction, achieving both high accuracy and generalizability. By blending algorithms with diverse learning mechanisms and inductive biases, the stacking model achieved superior results across all metrics, outperforming each individual learner. These findings highlight the importance of ensemble learning frameworks in psychiatric modeling. Figure 4 illustrates the comparative performance of the models using bar plots of evaluation metrics such as MSE, RMSE, MAE, and R^2. Additionally, Table 3 summarizes algorithm-specific strengths and limitations in predicting depression severity. The consistency of ensemble models across subgroups and variable types suggests promising applications in both clinical screening and public health surveillance.

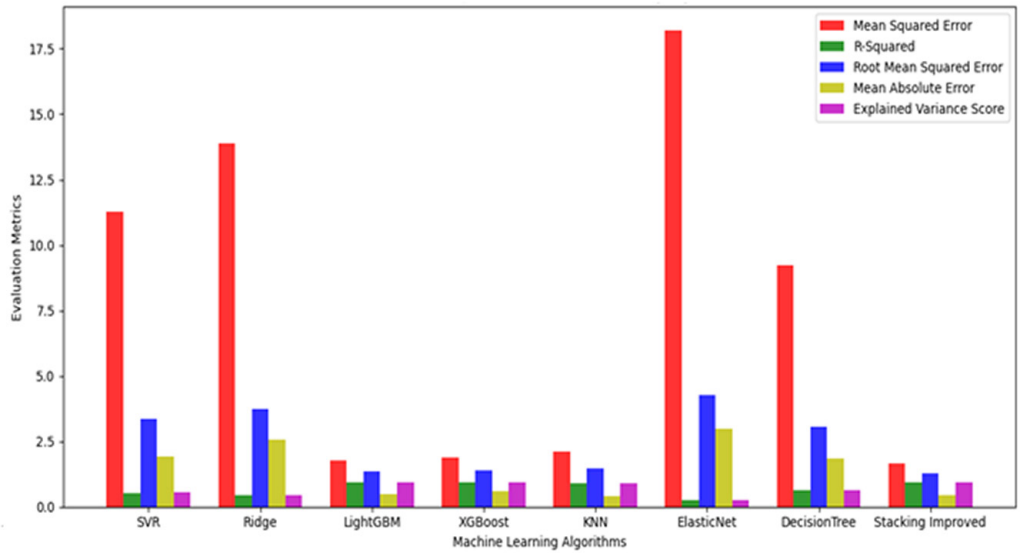

Fig. 4. Performance comparison of each base learner and the stacked ensemble

Table 3. Summary of model strengths and limitations in predicting depression severity

Model	Strengths	Limitations	Relevance to Depression Prediction
SVR	Captures complex, non-linear relationships; robust to overfitting	Computationally intensive; sensitive to kernel and parameter settings	Useful for modeling nuanced biomedical–psychological interactions
Ridge Regression	Simple, fast, and interpretable; handles multicollinearity in dense data	Limited ability to capture non-linear and interactive effects	Best for establishing linear baselines with lifestyle data
LightGBM	Efficient with large, sparse, and high-dimensional datasets; high accuracy	Requires careful tuning; reduced interpretability	Excels in capturing feature interactions among diverse biomedical markers
XGBoost	Powerful for structured data; handles missing values and complex patterns	Higher training time; prone to overfitting if not regulated	Strong fit for feature-rich mental health prediction tasks
KNN	Non-parametric; effective in identifying local data structures	Poor scalability with large datasets; sensitive to noise and irrelevant features	Adds value when patient similarity (e.g., demographics, symptoms) is relevant
ElasticNet	Performs automatic feature selection; handles correlated predictors	Struggles with non-linear dependencies; low predictive performance in this study	Limited value as a standalone model; helpful as a regularized ensemble input
Decision Tree	Easy to interpret; can capture simple non-linear splits	High variance; unstable without pruning or ensemble integration	Suitable as a component in stacking frameworks for early-stage testing
Stacking Regressor	Leverages strengths of diverse models; highly generalizable and accurate	Computational complexity requires careful validation and tuning	Best-performing architecture; ideal for high-dimensional integrated datasets

3.3 Significance for theory and practice

Theoretically, heterogeneous stacking improves variance explanation for continuous PHQ-9 targets. Practically, integrating biomedical and lifestyle data yields a deployable pipeline for stepped-care triage and monitoring workflows in health systems.

4 CONCLUSION

This study presents a machine learning framework for predicting depression severity based on PHQ-9 scores, utilizing a stacked heterogeneous ensemble applied to integrated biomedical and lifestyle data from the 2013–2014 NHANES dataset. Drawing on six multimodal data modules—including demographic, dietary, examination, laboratory, medication, and questionnaire components. The framework enables a holistic analysis of biopsychosocial determinants of mental health. Through a combination of data fusion, rigorous preprocessing, and hybrid feature selection, the study constructed a two-stage ensemble learning pipeline consisting of seven base learners, including ElasticNet, Ridge, SVR, KNN, Decision Tree, LightGBM, and XGBoost, stacked beneath a Voting Regressor meta-model. This heterogeneous architecture was further optimized via stratified sampling

and inverse-frequency weighting to enhance sensitivity to underrepresented high-severity depression cases. Empirical evaluation highlights the superiority of this ensemble approach. The stacked model achieved a MSE of 1.66, an RMSE of 1.29, and an R^2 of 0.93, outperforming all individual learners, including top performers such as LightGBM (MSE = 1.78) and XGBoost (MSE = 1.88). These results affirm that stacking not only enhances predictive accuracy but also improves resilience across variable types and patient subgroups by leveraging the diverse inductive biases of its constituent models.

The findings further reveal that linear models (e.g., Ridge, ElasticNet) underperform in capturing the non-linear, interactive dynamics of mental health data, while KNN and SVR models offer meaningful contributions when embedded within a more complex ensemble. Importantly, the framework's design ensures interpretability through tree-based components and systematic hyperparameter tuning, paving the way for explainable, clinician-friendly tools in precision psychiatry. In conclusion, the Stacked Heterogeneous Learning model developed in this study represents a scalable, accurate, and clinically relevant solution for predicting depression severity. It demonstrates how combining model diversity, data richness, and algorithmic rigor can yield powerful insights into mental health conditions.

5 FUTURE WORK

Future research should strengthen evaluation by reporting 95% confidence intervals for regression metrics (MSE, RMSE, MAE, R^2, Adjusted R^2, EVS) via bootstrap resampling while conducting paired statistical tests against strong baselines such as LightGBM. Adjusted R^2, defined as $1 - (1 - R^2)(n - 1)/(n - p - 1)$, should be systematically reported to account for model complexity, and robustness further examined through nested cross-validation or validation on independent cohorts. Methodological refinements may also include advanced imputation strategies (e.g., KNN or multivariate iterative imputation). Beyond accuracy, emphasis should shift toward clinical interpretability and deployment, with outputs translated into actionable insights such as symptom-specific risk profiling, supported by longitudinal studies leveraging mobile and wearable devices to capture real-world trajectories, the inclusion of longitudinal behavioral data, genetic profiles, and wearable-derived indicators to further enhance predictive performance, and enable just-in-time interventions. Finally, ethical compliance and data transparency must remain central, ensuring privacy, informed consent, equitable subgroup representation, and reproducibility through secure data handling and open methodological reporting.

6 REFERENCES

[1] S. Crowe, A. F. Howard, and B. Vanderspank, "The mental health impact of the COVID-19 pandemic on Canadian critical care nurses," *Intensive & Critical Care Nursing*, vol. 71, p. 103241, 2022. https://doi.org/10.1016/j.iccn.2022.103241

[2] F. Cosci, K. S. Christensen, S. Ceccatelli, C. Patierno, and D. Carrozzino, "Patient health questionnaire-9: A clinimetric analysis," *Braz. J. Psychiatry*, vol. 46, pp. 1–6, 2024. https://doi.org/10.47626/1516-4446-2023-3449

[3] US Preventive Services Task Force, "Screening for depression and suicide risk in adults: US Preventive Services Task Force recommendation statement," *JAMA*, vol. 329, no. 23, pp. 2057–2067, 2023. https://doi.org/10.1001/jama.2023.9297

[4] S. M. Ajibade *et al.*, "A literature review of machine learning techniques for recommender systems in e-commerce," in *Proc. 2024 IEEE 5th Int. Conf. Electro-Comput. Technol. Humanity (NIGERCON)*, 2024, pp. 1–6. https://doi.org/10.1109/NIGERCON62786.2024.10927121

[5] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, "Predicting changes in depression severity using the PSYCHE-D model involving person-generated health data: Longitudinal case-control observational study," *JMIR Mhealth Uhealth*, vol. 10, no. 3, p. e34148, 2022. https://doi.org/10.2196/34148

[6] P. Chondros *et al.*, "Development of a prognostic model for predicting depression severity in adult primary patients with depressive symptoms using the diamond longitudinal study," *J. Affect. Disord.*, vol. 227, pp. 854–860, 2018. https://doi.org/10.1016/j.jad.2017.11.042

[7] S. Akbarova *et al.*, "Improving depression severity prediction from passive sensing: Symptom-profiling approach," *Sensors*, vol. 23, no. 1, p. 8866, 2023. https://doi.org/10.3390/s23218866

[8] J. Masih and W. Verbeke, "Immune system function and its relation to depression: How exercise can alter the immune system-depression dynamics," *J. Depress. Anxiety*, vol. 7, no. 4, pp. 1–7, 2018. https://doi.org/10.4172/2167-1044.1000325

[9] I. C. Obagbuwa, S. Danster, and O. C. Chibaya, "Supervised machine learning models for depression sentiment analysis," *Frontiers in Artificial Intelligence*, vol. 6, p. 1230649, 2023. https://doi.org/10.3389/frai.2023.1230649

[10] Z. Sabouri, N. Gherabi, M. Nasri, M. Amnai, H. El Massari, and I. Moustati, "Prediction of depression via supervised learning models: Performance comparison and analysis," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 19, no. 9, pp. 93–107, 2023. https://doi.org/10.3991/ijoe.v19i09.39823

[11] M. R. Kumar, K. Pooja, M. Udathu, J. L. Prasanna, and C. Santhosh, "Detection of depression using machine learning algorithms," *Int. J. Online Biomed. Eng. (iJOE)*, vol. 18, no. 4, pp. 155–163, 2022. https://doi.org/10.3991/ijoe.v18i04.29051

[12] K. Ueafuea *et al.*, "Potential applications of mobile and wearable devices for psychological support during the COVID-19 pandemic: A review," *IEEE Sens. J.*, vol. 21, no. 6, pp. 7162–7178, 2021. https://doi.org/10.1109/JSEN.2020.3046259

[13] K. Vayadande, A. Bodhankar, A. Mahajan, and D. Prasad, "Classification of depression on social media using distant supervision," in *ITM Web of Conferences*, vol. 50, 2022. https://doi.org/10.1051/itmconf/20225001005

[14] Y. Sun, Z. Fu, Q. Bo, Z. Mao, X. Ma, and C. Wang, "The reliability and validity of PHQ-9 in patients with major depressive disorder in psychiatric hospital," *BMC Psychiatry*, 2020. https://doi.org/10.21203/rs.2.18098/v1

[15] S. Zulfiker, N. Kabir, A. Amin, and T. Nazneen, "An in-depth analysis of machine learning approaches to predict depression," *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, 2021. https://doi.org/10.1016/j.crbeha.2021.100044

[16] C. Lee and H. Kim, "Machine learning-based predictive modeling of depression in hypertensive populations," *PLoS ONE*, vol. 17, no. 7, p. e0272330, 2022. https://doi.org/10.1371/journal.pone.0272330

[17] A. Abd-Alrazaq, R. AlSaad, F. Shuweihdi, A. Ahmed, S. Aziz, and J. Sheikh, "Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression," *NPJ Digit. Med.*, vol. 6, pp. 1–16, 2023. https://doi.org/10.1038/s41746-023-00828-5

7 AUTHORS

Adefemi Ayodele is with the School of Architecture Computing and Engineering, University of East London, United Kingdom.

Adedotun Adetunla is with the Department of Mechanical Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa (E-mail: dotunadetunla@gmail.com).

Esther Akinlabi is with the Department of Mechanical Engineering, Colorado State University, Fort Collins, CO, USA.