JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

DEPRESSION PREDICTION USING STACKING MACHINE LEARNING MODEL

[1]B Suneetha, [2]Mrs V Dakshayani

[1]P G Student, [2] Associate Professor
[1] Dept of CSE, MJR College of Engineering and Technology, Piler, India
[2] Dept of CSE, MJR College of Engineering and Technology, Piler, India

Abstract: Depression is a major mental health concern often underdiagnosed due to subjective assessments and social stigma. Traditional diagnostic methods relying on clinical interviews and surveys are time-consuming and prone to bias. To overcome these limitations, this study proposes a Stacking Machine Learning Model that integrates SVM, KNN, and LightGBM as base learners, with Logistic Regression as the meta-classifier. The dataset includes socio-economic and lifestyle factors such as age, gender, marital status, income, and living expenses. After preprocessing using missing value handling, normalization, and feature correlation analysis, the stacking ensemble effectively combines the strengths of individual models—capturing nonlinear relations, locality-based learning, and complex feature interactions. Experimental results show that the proposed model outperforms individual classifiers in accuracy, precision, recall, and ROC-AUC, providing an efficient tool for early depression prediction and supporting data-driven mental health interventions.

I. INTRODUCTION

Depression is a widespread mental health disorder that affects millions of people worldwide, leading to emotional distress and reduced quality of life. According to the World Health Organization (WHO), over 280 million individuals suffer from depression, making it one of the leading causes of disability. Despite its prevalence, early detection remains difficult due to overlapping symptoms, subjective assessments, and social stigma. Traditional diagnostic methods based on clinical interviews and self-reported surveys are often time-consuming, inconsistent, and prone to bias. Hence, there is a growing need for automated, data-driven systems that can assist in accurate and early depression prediction. Machine Learning (ML) offers promising solutions by analyzing socio-economic, behavioral, and lifestyle data to identify depression patterns. However, single classifiers such as SVM, KNN, or Gradient Boosting often struggle with data imbalance, noise, and limited generalization. To address these issues, this research proposes a Stacking Ensemble Model combining SVM, KNN, and LightGBM as base learners, with Logistic Regression as the meta-classifier. The model aims to improve prediction accuracy, robustness, and interpretability, offering a reliable tool for early depression detection and supporting mental health professionals in timely intervention.

II. LITERATURE SURVEY

Several studies have explored stress and depression detection using physiological, behavioral, and machine learning-based approaches. *Jinlong Chao et al.* utilized functional near-infrared spectroscopy (fNIRS) to distinguish patients with Major Depressive Disorder (MDD) from healthy controls, achieving up to **99.94% accuracy**, and identified neuromarkers in the prefrontal cortex as indicators of depression. *Zhang et al*. examined the effects of acute psychosocial stress on cooperation and competition in women using fNIRS, revealing that stress enhanced competitive behavior and neural synchronization in the dorsolateral prefrontal cortex. *Soyeon Park and Suh-Yeon Dong* improved mental state classification by incorporating daily stress levels into fNIRS-based models, demonstrating that personalized stress assessment enhances detection accuracy. *Jayawickrama and Rupasingha* investigated stress detection through sleep habits using multiple machine learning algorithms, where the **Naïve Bayes** classifier achieved **91.27% accuracy**. *Nina Speicher et al.* analyzed how acute stress, gender, and personality traits affect moral decision-making, concluding that stress promotes prosocial tendencies, especially in females. *V.R. Archana and B.M. Devaraju* predicted stress levels using physiological signals like ECG, EMG, GSR, and respiration, showing improved accuracy with models such as Decision Tree, Naïve Bayes, and KNN. Similarly, *R. Swarna Malika and I. Ravi* compared Logistic Regression, Decision Tree, and Random Forest algorithms on a stress dataset, identifying **Random Forest** as the most accurate classifier. These studies collectively highlight the growing role of AI and physiological data in enhancing mental health prediction and stress assessment accuracy.

III. PROPOSED METHODOLOGY

3.1 Problem Statement

Traditional depression prediction systems rely mainly on clinical interviews, self-assessment surveys such as PHQ-9 and BDI, and single-model machine learning techniques like SVM, KNN, and Random Forest. While these methods help identify depression patterns, they often suffer from subjectivity, inconsistency, and limited scalability. Machine learning-based approaches have

improved automation but still face challenges such as poor generalization, sensitivity to noise, and data imbalance—where non-depressed cases dominate the dataset, causing biased predictions. Single classifiers also fail to perform consistently across heterogeneous socio-economic data. Hence, a hybrid ensemble approach is required to enhance accuracy, robustness, and interpretability in depression prediction.

3.2 Proposed System

The proposed system introduces a Stacking Machine Learning Framework that integrates multiple classifiers for improved prediction performance. The framework combines Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Light Gradient Boosting Machine (LightGBM) as base learners, with Logistic Regression (LR) as the meta-classifier. This ensemble leverages the complementary strengths of individual algorithms to produce more reliable predictions. The methodology involves four main stages: data preprocessing, model construction, training/testing, and performance evaluation. In preprocessing, missing values are imputed using mean or median techniques, and data normalization is applied using RobustScaler to minimize the influence of outliers. The model construction phase employs SVM for nonlinear separation, KNN for distance-based learning, and LightGBM for efficient feature interaction modeling. Predictions from these models are combined through a Logistic Regression meta-learner to generate the final classification. During training, stratified sampling ensures balanced class representation, and cross-validation prevents overfitting. Performance evaluation uses metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC, supported by confusion matrices and ROC curves for visualization. The proposed ensemble provides a more stable and accurate prediction framework than single learners by handling nonlinearity, noise, and imbalance more effectively. It also offers interpretability through the logistic meta-model, making it suitable for mental health professionals to identify at-risk individuals early.

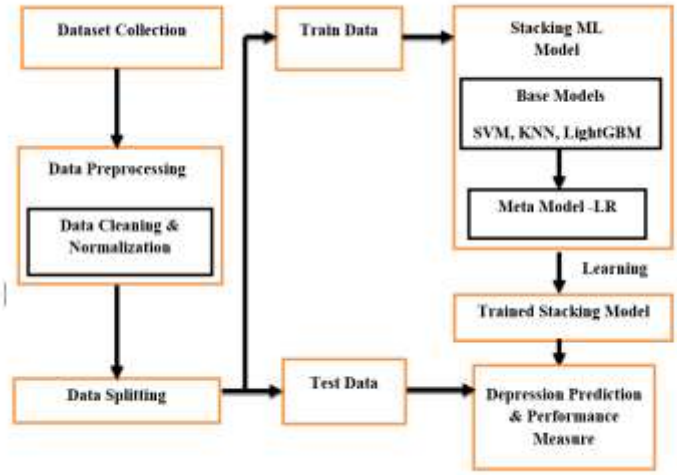


FIG 1: SYSTEM ARCHITECTURE

IV. SOFTWARE AND DOMAIN DESCRIPTION

Python is one of the most popular programming languages used in machine learning and data science due to its simplicity and extensive library support. Commonly used libraries include NumPy for numerical computations, SciPy for optimization and statistics, Scikit-learn for machine learning algorithms, Theano for mathematical expression optimization, TensorFlow and PyTorch for deep learning and neural networks, and Keras for easy and fast neural network prototyping. Pandas aids in data extraction, cleaning, and analysis, while Matplotlib and Seaborn are used for effective data visualization. OpenCV supports real-time computer vision applications, Pillow enables image manipulation, and ImageAI provides APIs for object detection using pre-trained models. Together, these libraries form a robust ecosystem for developing intelligent and data-driven systems.

Data mining, an interdisciplinary field integrating databases, statistics, machine learning, and visualization, focuses on extracting meaningful patterns and hidden insights from large datasets. It involves processes such as data preprocessing, transformation, and modeling to uncover relationships and trends useful for decision-making. Techniques like classification, association, and sequence analysis help identify behavioral patterns, correlations, and time-based trends in data. Data warehouses and OLAP tools further support efficient storage, retrieval, and multi-dimensional data analysis. Overall, data mining and machine learning together provide powerful tools for predictive analytics, helping in areas such as business intelligence, healthcare, and social sciences by transforming raw data into actionable knowledge.

SYSTEM DESIGN

System design defines how data flows, processes, and outputs are represented within a system. The Data Flow Diagram (DFD) illustrates input, processing, and output relationships, showing how data moves through various transformations. The Unified Modeling Language (UML) provides standardized diagrams such as Use Case, Class, Sequence, and Activity diagrams to model system behavior, structure, and workflows. Input design focuses on creating user-friendly, error-free, and secure data entry methods, while output design ensures that processed information is presented clearly and effectively to support decision-making. System

testing verifies that the integrated software meets all functional and user requirements through various levels, including unit, integration, functional, and acceptance testing. Both white box and black box testing approaches are used to ensure reliability, accuracy, and performance. All tests produced successful results, confirming the proper functionality of the system.

V. EXPERIMENTAL SYSTEMS

MODULE 1- Data Loading and EDA

Dataset

	Survey_id	Ville_id	sex	Age	Married	Number_children	education_level	total_members	gained_asset	durable_asset	...	incomeg_salary	incoming_own_farm
0	926	91	1	28	1	4	10	5	28912201	22861940	...	0	0
1	747	57	1	23	1	3	8	5	28912201	22861940	...	0	0
2	1196	115	1	22	1	3	9	5	28912201	22861940	...	0	0
3	1065	97	1	27	1	2	10	4	52667106	19686864	...	0	1
4	806	42	0	59	0	4	10	6	82606267	17352654	...	1	0
5	483	25	1	35	1	6	10	8	35937466	736707	...	0	1
6	849	130	0	34	0	1	9	3	41303144	21925041	...	0	0
7	500	195	1	32	1	7	9	9	11087568	2522428	...	1	0
8	380	33	1	29	1	4	10	5	28912201	22861940	...	0	0
9	540	52	1	84	0	0	1	5	28912201	22861940	...	0	0

Dataset shape / Dataset input features / data types / Statistical Analysis

```
df.shape
```

```
(1767, 23)
```

```
df.columns
```

```
Index(['Survey_id', 'Ville_id', 'sex', 'Age', 'Married', 'Number_children',
       'education_level', 'total_members', 'gained_asset', 'durable_asset',
       'save_asset', 'living_expenses', 'other_expenses', 'incoming_salary',
       'incoming_own_farm', 'incoming_business', 'incoming_no_business',
       'incoming_agricultural', 'farm_expenses', 'labor_primary',
       'lasting_investment', 'no_lasting_investmen', 'depressed'],
      dtype='object')
```

```
RangeIndex: 1767 entries, 0 to 1766
Data columns (total 23 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Survey_id             1767 non-null   int64
 1   Ville_id              1767 non-null   int64
 2   sex                   1767 non-null   int64
 3   Age                   1767 non-null   int64
 4   Married               1767 non-null   int64
 5   Number_children       1767 non-null   int64
 6   education_level       1767 non-null   int64
 7   total_members         1767 non-null   int64
 8   gained_asset          1767 non-null   int64
 9   durable_asset         1767 non-null   int64
 10  save_asset            1767 non-null   int64
 11  living_expenses       1767 non-null   int64
 12  other_expenses        1767 non-null   int64
 13  incoming_salary       1767 non-null   int64
 14  incoming_own_farm     1767 non-null   int64
 15  incoming_business     1767 non-null   int64
 16  incoming_no_business  1767 non-null   int64
 17  incoming_agricultural 1767 non-null   int64
 18  farm_expenses         1767 non-null   int64
 19  labor_primary         1767 non-null   int64
 20  lasting_investment    1767 non-null   int64
 21  no_lasting_investmen  1743 non-null   float64
 22  depressed             1767 non-null   int64
dtypes: float64(1), int64(22)
memory usage: 317.6 KB
```

index	count	mean	std	min	25%	50%	75%	max
Survey_id	1767.0	715.9501980758347	413.1982600221414	1.0	358.5	712.0	1077.5	1429.0
Ville_id	1767.0	77.54612337284885	67.31296385140399	1.0	24.0	58.0	107.5	292.0
sex	1767.0	0.9151103395385025	0.2787993059776255	0.0	1.0	1.0	1.0	1.0
Age	1767.0	35.44821731748727	14.49608475833424	17.0	25.0	31.0	43.0	91.0
Married	1767.0	0.7617771383883605	0.4264840824933954	0.0	1.0	1.0	1.0	1.0
Number_children	1767.0	2.884500084888433	1.9019842075301113	0.0	2.0	3.0	4.0	11.0
education_level	1767.0	8.559139784946236	3.0458526599822436	1.0	7.5	9.0	10.0	19.0
total_members	1767.0	5.001697792869027	1.7783456939822351	1.0	4.0	5.0	6.0	12.0
gained_asset	1767.0	33554721.49632145	19846281.244199434	325112.0	24130047.0	28912231.0	37172832.0	99127548.0
durable_asset	1767.0	27507449.85681947	10463244.300703166	162556.0	19386808.0	22861949.0	29305823.0	89615601.0
save_asset	1767.0	27586600.88189021	10109962.45355671	172966.0	23399979.0	23399979.0	23399979.0	89826758.0
living_expenses	1767.0	32199489.17600453	20856626.18119615	262919.0	20419598.0	28692283.0	39835351.0	89295282.0
other_expenses	1767.0	33788313.49558517	21015555.629682555	172966.0	21412550.0	28203088.0	40589863.0	89623799.0
incoming_salary	1767.0	0.17940011318619128	0.3837056144624654	0.0	0.0	0.0	0.0	1.0
incoming_own_farm	1767.0	0.25297113775212224	0.4348379127350179	0.0	0.0	0.0	1.0	1.0
incoming_business	1767.0	0.10243350311262026	0.3033033831108123	0.0	0.0	0.0	0.0	1.0
incoming_no_business	1767.0	0.2558007923033338	0.4364338823597455	0.0	0.0	0.0	1.0	1.0
incoming_agricultural	1767.0	34399866.70967742	20715897.040546328	325112.0	22955363.5	30028816.0	40038424.0	89789095.0
farm_expenses	1767.0	35426469.03956027	21079866.38859488	271505.0	22810781.0	31363432.0	42983455.0	99651194.0
labor_primary	1767.0	0.21165817778232032	0.4085594007589239 5	0.0	0.0	0.0	0.0	1.0
lasting_investment	1767.0	32910621.524052065	21248231.199852866	74292.0	20019113.0	28411718.0	39424827.5	88446667.0
no_lasting_investmen	1743.0	34671181.063186584	21848259.90939139	126312.0	21220384.5	28292797.0	42108401.0	99651194.0
depressed	1767.0	0.3259762308898324	0.4688711359532240 5	0.0	0.0	0.0	1.0	1.0

MODULE 2 - Data Preprocessing

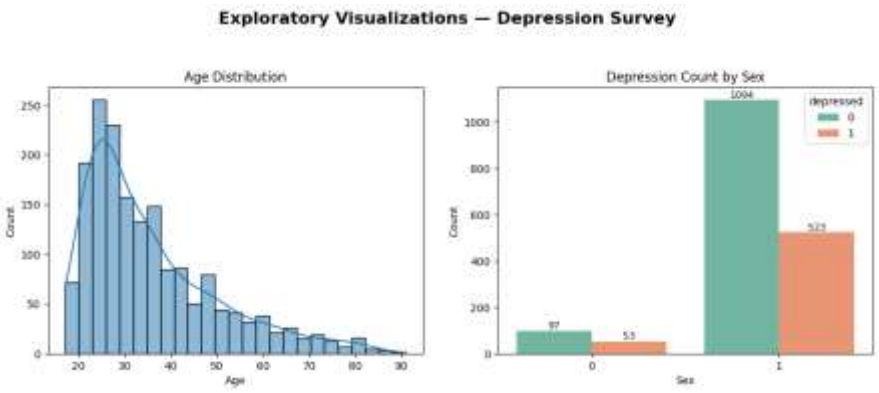
Data Cleaning

Null Value Checking Result	After Cleaning
Survey_id 0 Ville_id 0 sex 0 Age 0 Married 0 Number_children 0 education_level 0 total_members 0 gained_asset 0 durable_asset 0 save_asset 0 living_expenses 0 other_expenses 0 incoming_salary 0 incoming_own_farm 0 incoming_business 0 incoming_no_business 0 incoming_agricultural 0 farm_expenses 0 labor_primary 0 lasting_investment 0 no_lasting_investmen 24 depressed 0 dtype: int64	Survey_id 0 Ville_id 0 sex 0 Age 0 Married 0 Number_children 0 education_level 0 total_members 0 gained_asset 0 durable_asset 0 save_asset 0 living_expenses 0 other_expenses 0 incoming_salary 0 incoming_own_farm 0 incoming_business 0 incoming_no_business 0 incoming_agricultural 0 farm_expenses 0 labor_primary 0 lasting_investment 0 no_lasting_investmen 0 depressed 0 dtype: int64

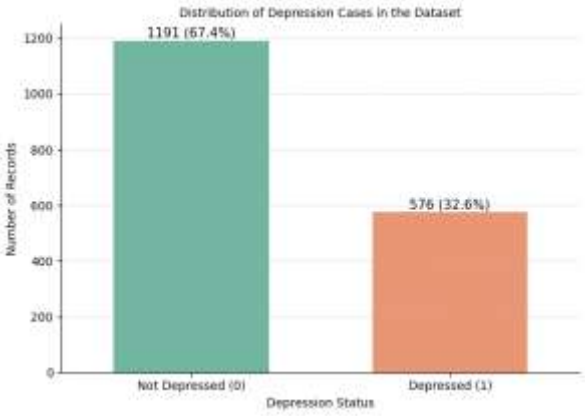
Data Scaling using Robust Scaler

```
array([[ 0.29763561,  0.39520958,  0.        , ...,  0.        ,
         0.        ,  0.        ],
       [ 0.04867872, -0.01197605,  0.        , ...,  0.        ,
         0.        ,  0.        ],
       [ 0.66481224,  0.68263473,  0.        , ...,  0.        ,
         0.        ,  0.        ],
       ...,
       [-0.86369958, -0.58682635,  0.        , ...,  0.        ,
        -0.51500515, -0.86717768],
       [-0.65890403, -0.10778443,  0.        , ...,  0.        ,
         3.85921068,  0.11393379],
       [-0.90125174, -0.5508982 ,  0.        , ...,  0.        ,
         0.73162057,  2.13145373]])
```

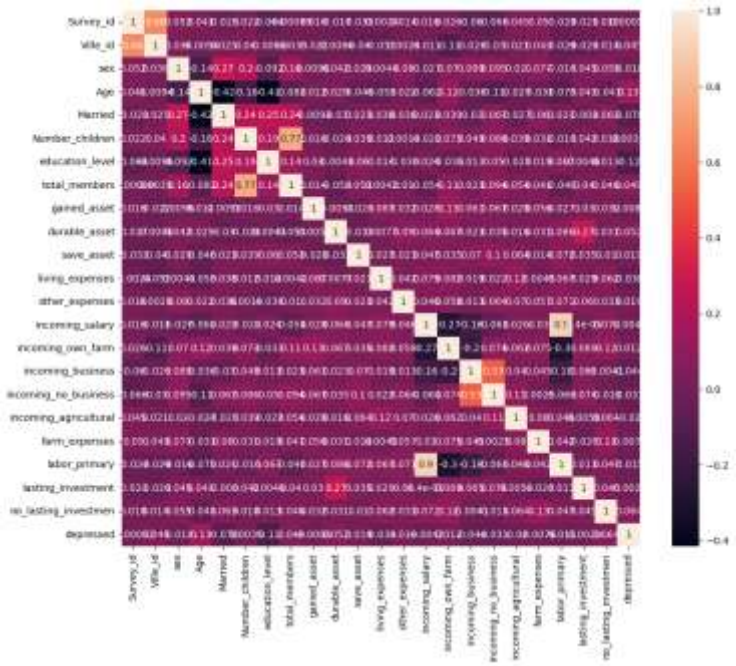
Data Visualization



Outcomes Distribution



Correlation between all the features

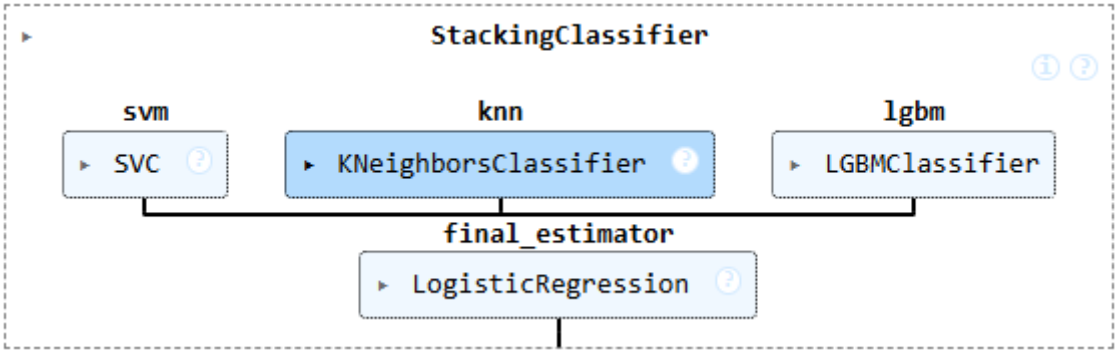


MODULE 3 – Data Splitting

Data Splitting – 80:20

```
training set size: (1413, 22), testing set size: (354, 22)
```

MODULE 4 & 5 – Stacking ML Model Building and Training



MODULE 6 – Depression Prediction

Testing Set Prediction Result

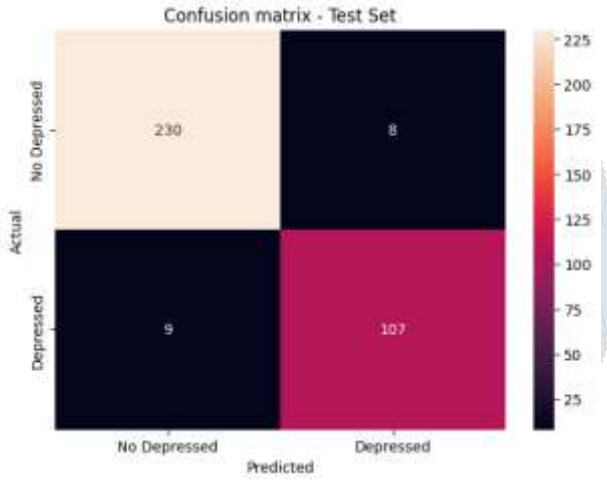
✅ Testing Set Prediction Results (first 20 rows):

Index	Actual Label	Predicted Label	Predicted Probability
0	1.0	1.0	0.861
1	1.0	1.0	0.849
2	0.0	0.0	0.099
3	0.0	0.0	0.084
4	0.0	0.0	0.07
5	0.0	0.0	0.188
6	0.0	0.0	0.072
7	0.0	0.0	0.173
8	0.0	0.0	0.087
9	1.0	1.0	0.867
10	0.0	0.0	0.184
11	0.0	0.0	0.088
12	0.0	0.0	0.075
13	1.0	1.0	0.864
14	0.0	0.0	0.226
15	1.0	1.0	0.867
16	0.0	0.0	0.391
17	1.0	1.0	0.864
18	0.0	0.0	0.134
19	0.0	0.0	0.07

The proposed Stacking Machine Learning Model was tested on an unseen dataset to evaluate its accuracy in predicting depression status. The model effectively distinguished between depressed and non-depressed individuals, as shown in the test results containing actual labels, predicted labels, and probability scores ranging from 0.07 to 0.86, indicating strong classification confidence. These results confirm that the model achieves high accuracy, stable probability estimates, and strong generalization, proving its reliability as a robust tool for early depression detection and mental health assessment.

MODULE 7 – Performance Evaluation

Confusion Matrix



The confusion matrix demonstrates the strong performance of the proposed Stacking Machine Learning Model in predicting depression. It accurately classified 230 “No Depressed” (True Negatives) and 107 “Depressed” (True Positives) cases, with only 8 False Positives and 9 False Negatives. This balanced and highly accurate performance indicates the model’s effectiveness in distinguishing between depressed and non-depressed individuals with minimal confusion. The integration of SVM, KNN, and LightGBM base learners, combined with Logistic Regression as a meta-learner, enables efficient handling of nonlinear relationships and subtle data variations. Overall, the results confirm the model’s robustness, precision, and suitability for reliable mental health prediction and early detection.

Performance Metrics

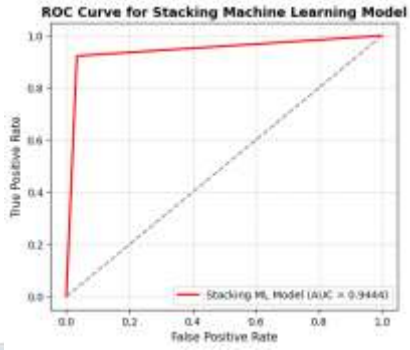
Accuracy on test set: 95.19999999999999%

	precision	recall	f1-score	support
0.0	0.9623	0.9664	0.9644	238
1.0	0.9304	0.9224	0.9264	116
accuracy			0.9520	354
macro avg	0.9464	0.9444	0.9454	354
weighted avg	0.9519	0.9520	0.9519	354

The proposed Stacking Machine Learning Model achieved excellent performance in depression prediction, with an overall accuracy of 95.20%. Evaluation using Accuracy, Precision, Recall, and F1-Score confirmed its strong and balanced classification

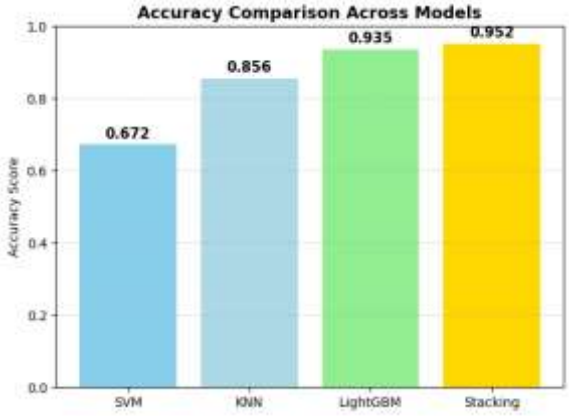
ability. For the "No Depression" class, it achieved precision, recall, and F1-score values of 0.9623, 0.9664, and 0.9644, while for the "Depression" class, the values were 0.9304, 0.9224, and 0.9264, respectively. The macro and weighted F1-scores (0.9454 and 0.9519) further validate consistent performance across both classes. By integrating SVM, KNN, and LightGBM with Logistic Regression as a meta-learner, the stacking ensemble effectively combines nonlinear decision boundaries, local pattern recognition, and feature interaction learning. Overall, the model demonstrates high accuracy, robustness, and reliability, making it suitable for early depression detection and mental health prediction.

ROC Curve



The ROC curve of the proposed Stacking Machine Learning Model demonstrates excellent discrimination between depressed and non-depressed individuals, with an AUC value of 0.9444. The curve's steep rise toward the top-left corner indicates a high true positive rate and low false positive rate, confirming strong sensitivity and specificity. By combining SVM, KNN, and LightGBM through Logistic Regression as a meta-learner, the model effectively handles nonlinear patterns and complex feature interactions. Overall, the high AUC score validates the model's strong predictive power, reliability, and suitability for early depression detection and AI-assisted mental health diagnosis.

Performance Comparison



The accuracy comparison chart shows that the proposed Stacking Machine Learning Model outperforms individual models—SVM, KNN, and LightGBM—in depression prediction. SVM achieved 67.2% accuracy, KNN reached 85.6%, and LightGBM improved to 93.5%, while the stacking ensemble attained the highest accuracy of 95.2%. By combining SVM, KNN, and LightGBM through Logistic Regression as a meta-classifier, the stacking model effectively merges their strengths, achieving better generalization and reducing errors. This demonstrates its superior reliability, robustness, and suitability for real-world depression detection compared to standalone models.

VI. CONCLUSION

Early detection of depression is essential to prevent severe mental health issues, but traditional methods relying on self-reports are often unreliable. To address this, a Stacking Machine Learning Model integrating SVM, KNN, and LightGBM with Logistic Regression as a meta-learner was developed to predict depression using socio-economic and lifestyle data. After preprocessing with missing value handling and normalization, the proposed model achieved a 95.2% accuracy, outperforming individual models (SVM: 67.2%, KNN: 85.6%, LightGBM: 93.5%). With high precision, recall, F1-score, and an AUC of 0.9444, the stacking ensemble demonstrated strong discriminative ability and robust performance. This approach proves that ensemble learning enhances depression prediction accuracy and reliability, offering a data-driven tool for early diagnosis and intervention. Future work may integrate deep learning and explainable AI for improved interpretability and clinical usability.

REFERENCES

[1] V. R. Archana and B. M. Devaraju, “Stress Detection Using Machine Learning Algorithms”, IJRESM, vol. 3, no. 8, pp. 251–256, Aug. 2020.

[2] G. Jayawickrama, R.A.H.M. Rupasingha, “Human Stress Detection Based on Sleeping Habits Using Machine Learning Algorithms”, 2023 3rd International Conference on Advanced Research in Computing (ICARC), IEEE Conference, 2023.

[3] R. Swarna Malika, I. Ravi, "Stress Detection Using Machine Learning Techniques", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.10, Issue 3, page no.54-58, March-2023,

[4] V, Sudha and S, Kaviya and S, Sarika and R, Raja, Stress Detection Based on Human Parameters Using Machine Learning Algorithms (February 10, 2022). Proceedings of the International Conference on Innovative Computing & Communication (ICICC) 2022.

[5] Md Minhazur Rahman, ASM Mohaimenul Islam, Jonayet Miah, “sleepWell: Stress Level Prediction Through Sleep Data. Are You Stressed?”

[6] Jinlong Chao, Shuzhen Zheng, Hongtong Wu, Dixin Wang, Xuan Zhang, Hong Peng, Bin Hu:” fNIRS Evidence for Distinguishing Patients With Major Depression and Healthy Controls:” Date of publiucation on September 2021.

[7] Zhang, R., Zhou, X., Feng, D., Yuan, D., Li, S., Lu, C., & Li, X. (2021). Effects of acute psychosocial stress on interpersonal cooperation and competition in young women. *Brain and Cognition*, *151*.

[8] Jerome Brunelin, Shirley Fecteau.” Impact of bifrontal transcranial Direct Current Stimulation on decision-making and stress reactivity. A pilot study.” Journal of Psychiatric Research 135(5). Date of publication December 2020.

[9] Soyeon Park, Suh-Yeon Dong.” Effects of Daily Stress in Mental State Classification.” IEEE Access. License CC BY 4.0.Published on 2020.

[10] Nina Speicher, Monika Sommer, Stefan Wüst.” Effects of gender and personality on everyday moral decision-making after acute stress exposure.” Psychoneuroendocrinology 124(67).Date of publication December 2020.

[11] Li, R., Liu, Z. Stress detection using deep neural networks. BMC Med Inform Decis Mak 20 (Suppl 11), 285 (2020).

[12] Ahuja, R. and Banga, A., 2019. Mental stress detection in university students using machine learning algorithms. Procedia Computer Science, 152, pp.349-353.