IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 7, July 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

Obesity Prediction using Support Vector Machine: A Machine Learning Approach for Health Risk Assessment

Soumya K S[1], Alias Itten[2]

Lecturer, Department of Computer Engineering, Govt. Polytechnic College, Kalamassery, Ernakulam, Kerala, India[1]

Lecturer, Department of Computer Engineering, Govt. Polytechnic College, Kaduthuruthy, Kottayam, Kerala, India[2]

ABSTRACT: Obesity is a significant global health concern associated with chronic diseases such as diabetes, cardiovascular disorders, and hypertension. Early detection and classification of obesity levels can help in preventive healthcare and personalized interventions. This study explores the effectiveness of Support Vector Machine (SVM) classifiers in predicting obesity categories based on lifestyle and physiological attributes. The Kaggle **Obesity Prediction Dataset** is utilized, containing demographic, dietary, and physical activity-related features. Data preprocessing techniques, including label encoding for categorical variables and feature standardization, are applied to enhance model performance. The dataset is split into training and testing subsets, and an SVM classifier with an **RBF kernel** is trained for classification. Model performance is evaluated using **accuracy, precision, recall, F1-score, and a confusion matrix**. The results demonstrate the effectiveness of SVM in classifying obesity levels with high accuracy. Additionally, **permutation importance** is applied to analyze the contribution of individual features to obesity prediction. The confusion matrix and visualizations highlight misclassification patterns, providing insights into potential improvements. The study suggests that machine learning techniques, particularly **SVM-based classification**, can play a crucial role in obesity risk assessment. Future research can extend this work by integrating deep learning models or hybrid approaches to improve prediction accuracy further. This study contributes to the growing field of **AI-driven healthcare solutions** and emphasizes the importance of data-driven decision-making in public health.

KEYWORDS: Obesity Prediction, Support Vector Machine (SVM), Machine Learning in Healthcare, Classification Models, Feature Importance Analysis, Confusion Matrix, Health Risk Assessment, Data-Driven Decision Making.

I. INTRODUCTION

Obesity is a growing global health challenge that significantly increases the risk of chronic diseases such as diabetes, cardiovascular disorders, and hypertension. The World Health Organization (WHO) classifies obesity as an epidemic due to its rising prevalence across different age groups and geographical regions. Several factors contribute to obesity, including genetic predisposition, dietary habits, physical activity levels, and socioeconomic conditions. Traditional methods for assessing obesity rely on Body Mass Index (BMI) and clinical evaluations, which may not always provide an accurate classification of an individual's health risks. With the rise of artificial intelligence and machine learning, predictive models have become valuable tools for assessing obesity levels based on lifestyle and physiological attributes. Machine learning algorithms can process large datasets, identify hidden patterns, and make reliable predictions, offering a more data-driven approach to obesity classification.

Among various machine learning techniques, **Support Vector Machines (SVMs)** have shown high accuracy and robustness in classification problems. SVM is particularly useful for handling complex, nonlinear relationships between features and target variables. This study employs an SVM classifier to predict obesity levels using the Kaggle **Obesity Prediction Dataset**, which includes multiple attributes such as eating habits, physical activity, and demographic information. The dataset undergoes preprocessing, including categorical encoding, feature scaling, and train-test splitting, to optimize the model's performance. The trained SVM model is evaluated using metrics such as **accuracy, precision, recall, F1-score, and a confusion matrix** to assess its effectiveness. Additionally, **permutation importance analysis** is conducted to determine the most influential features in predicting obesity. The findings of this study aim to demonstrate the applicability of SVM in healthcare analytics and provide insights into the role of machine

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

learning in obesity risk assessment. This research contributes to the growing field of **AI-driven healthcare solutions**, paving the way for automated, data-driven decision-making in medical diagnostics and public health planning.

II. LITERATURE SURVEY

Joshi et al. (2023)
This study applied various machine learning models, including Random Forest, XGBoost, Decision Tree, k-Nearest Neighbors, Support Vector Machine, Linear Regression, Naïve Bayes, and Multilayer Perceptron, to predict obesity. The models were evaluated on recall, accuracy, F1-score, and precision using both generalized and gender-segregated data. The findings provided insights into feature selection and early obesity identification, demonstrating the performance of algorithms on different datasets.

Lee et al. (2022)
The researchers developed a model predicting obesity risk by analyzing genome-wide and epigenome-wide interactions, focusing on dietary factors. Utilizing the generalized multifactor dimensionality reduction method, they identified significant predictors and applied machine learning algorithms, achieving an accuracy of 70% with a ROC-AUC of 0.72. Key predictors included specific SNPs, DNA methylation sites, and dietary factors like processed meat and diet soda consumption.

Gupta et al. (2022)
This study employed deep learning approaches with interpretable elements to predict obesity using electronic health record (EHR) data. The model demonstrated improved prediction accuracy and provided insights into the contributing factors to obesity, enhancing the interpretability of the results.

Rodríguez et al. (2021)
The authors applied machine learning techniques to predict overweight or obesity, focusing on feature selection and model performance. Their approach highlighted the importance of specific variables in predicting obesity, contributing to more targeted prevention strategies.

Mondal et al. (2023)
The study utilized machine learning classifiers to predict childhood obesity based on single and multiple well-child visit data. The models achieved high accuracy, demonstrating the potential of early prediction and intervention strategies in pediatric populations.

Thamrin et al. (2021)
This research analyzed Indonesian health data using machine learning techniques to predict adult obesity. The findings emphasized the role of socioeconomic and lifestyle factors in obesity prediction within the Indonesian context.

Yagin et al. (2023)
The authors estimated obesity levels using a trained neural network approach optimized by the Bayesian technique. Their model demonstrated improved accuracy and robustness in predicting obesity levels.

Ferdowsy et al. (2021)
This study applied a machine learning approach to predict obesity risk, focusing on behavioral sciences. The model identified key behavioral factors contributing to obesity, offering insights for targeted interventions.

Anisat et al. (2022)
The researchers developed an obesity prediction model using machine learning techniques, achieving significant accuracy in their predictions. The study highlighted the potential of machine learning in developing effective obesity prediction models.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

Author(s) & Year	Dataset Used	Machine Learning Models	Best Model Accuracy (%)	Key Findings
Joshi et al. (2023)	Kaggle Obesity Dataset	SVM, Random Forest, XGBoost, Decision Tree, k-NN, Naïve Bayes	89.2% (XGBoost)	XGBoost outperformed other models in gender-based obesity prediction.
Lee et al. (2022)	Genome-wide & Epigenome-wide Data	Generalized Multifactor Dimensionality Reduction (GMDR), SVM, Decision Trees	70% (GMDR)	Gene-diet interactions significantly impact obesity risk.
Gupta et al. (2022)	Electronic Health Records (EHR)	Deep Learning (Neural Networks)	92.5% (Neural Network)	Deep learning models provide superior predictive performance for obesity risk.
Rodríguez et al. (2021)	Public Health Dataset	Logistic Regression, SVM, Random Forest	85.7% (Random Forest)	Feature selection improves obesity classification accuracy.
Mondal et al. (2023)	Pediatric Health Records	Decision Trees, SVM, Neural Networks	88.6% (Neural Network)	Early prediction of childhood obesity is feasible with ML models.
Thamrin et al. (2021)	Indonesian Health Research Data	Random Forest, SVM, k-NN	83.1% (Random Forest)	Socioeconomic factors play a crucial role in obesity classification.
Yagin et al. (2023)	Medical Dataset	Bayesian-Optimized Neural Network	91.3% (Neural Network)	Bayesian optimization improves obesity classification performance.
Ferdowsy et al. (2021)	Behavioral Sciences Data	Decision Trees, SVM, Random Forest	86.2% (Random Forest)	Behavioral factors are key predictors of obesity.
Anisat et al. (2022)	Public Health Dataset	SVM, Decision Trees, Neural Networks	87.4% (Neural Network)	Machine learning can enhance obesity risk assessment models.

Table1: Summary of Research Papers on Obesity Prediction

III. METHODOLOGY

The methodology for obesity prediction using **Support Vector Machine (SVM)** involves several stages, including **data preprocessing, model selection, training, evaluation, and feature importance analysis**. The steps are systematically designed to ensure high accuracy and robustness in classification.

1. Dataset and Preprocessing

The **Kaggle Obesity Prediction Dataset** is used, which consists of demographic, lifestyle, and dietary factors influencing obesity levels. The dataset undergoes the following preprocessing steps:

1. **Handling Missing Values:** Any missing values are imputed using statistical methods like mean, median, or mode.
2. **Encoding Categorical Variables:** Non-numeric features (e.g., gender, eating habits) are converted into numerical representations using **One-Hot Encoding** or **Label Encoding**.
3. **Feature Scaling:** Since SVM is sensitive to feature magnitudes, the data is standardized using **Min-Max Scaling**:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4. **Train-Test Split:** The dataset is divided into **80% training** and **20% testing** to evaluate model performance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

2. Support Vector Machine (SVM) Classifier

Support Vector Machine (SVM) is chosen due to its ability to handle high-dimensional data and classify non-linearly separable cases using kernels. Given a dataset with feature vectors $\mathbf{x_i}$ and class labels $\mathbf{y_i}$ **(i = 1, 2, ..., n)**, SVM finds an optimal hyperplane that separates classes by maximizing the margin.

The decision function in SVM is given by:

$$f(x) = \omega^T b$$

where **ω** is the weight vector, **x** is the input feature vector, and **b** is the bias term.

For a **linearly separable dataset**, SVM optimizes the margin by solving:

$$\min_{\omega,b} \frac{1}{2} \|\omega\|^2$$

subject to:

$$y_i(\omega^T x_i + b) \geq 1, \forall i$$

For **non-linearly separable data**, SVM introduces a **kernel function** K(x,x′) that maps input features to a higher-dimensional space. The **Radial Basis Function (RBF) Kernel** is applied:

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right)$$

3. Model Training and Evaluation

1. **Hyperparameter Tuning:** The **Grid Search Cross-Validation (GridSearchCV)** technique is applied to optimize **CCC** (regularization parameter) and **γ\gammaγ** (kernel coefficient).
2. **Performance Metrics:** The trained SVM model is evaluated using the following metrics:

- **Accuracy**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision**:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall**:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-score**:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where **TP**, **TN**, **FP** and **FN** represent **True Positives, True Negatives, False Positives, and False Negatives** respectively.

4. Feature Importance Analysis

Since SVM does not inherently provide feature importance, **Permutation Importance** is used to analyze which features contribute most to obesity prediction. The importance of a feature **j** is measured by shuffling its values and observing the change in model performance:

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

$$\text{FeatureImportance (j)} = \text{Accuracy}_{\text{original}} - \text{Accuracy}_{\text{shuffled(j)}}$$

where **Accuracy$_{original}$** is the model's accuracy before shuffling, and **Accuracy$_{shuffled(j)}$** is after shuffling feature **j**.

5. Visualization and Interpretation

1. **Confusion Matrix Graph:** A heatmap is used to illustrate correct and incorrect predictions.
2. **Feature Importance Graph:** A bar plot displays the most significant features affecting obesity classification.

This methodology outlines a systematic approach to obesity prediction using **SVM with RBF kernel**. The combination of **feature scaling, hyperparameter tuning, and performance evaluation metrics** ensures a robust classification model. The results highlight the role of **machine learning in healthcare analytics**, offering insights into key predictors of obesity. Future research can extend this work by integrating **deep learning or ensemble learning** techniques for improved accuracy.

IV. EXPERIMENTAL RESULTS

The experimental evaluation of the **Support Vector Machine (SVM) classifier** for obesity prediction was performed on the **Kaggle Obesity Prediction Dataset**. The model achieved a high classification accuracy of **89.13%**, indicating its strong predictive performance.

1. Model Performance Metrics

The **classification report** summarizes the model's effectiveness in predicting different obesity levels. The evaluation metrics include **precision, recall, and F1-score**, which provide deeper insights into the classifier's performance:

Class Label	Precision	Recall	F1-Score	Support
0 (Normal Weight)	0.93	0.95	0.94	56
1 (Overweight Level 1)	0.75	0.84	0.79	62
2 (Overweight Level 2)	0.95	0.90	0.92	78
3 (Obesity Type I)	0.93	0.98	0.96	58
4 (Obesity Type II)	1.00	1.00	1.00	63
5 (Obesity Type III)	0.79	0.73	0.76	56
6 (Underweight)	0.87	0.82	0.85	50

Table2: Model Performance

- The **macro average** F1-score is **0.89**, showing balanced predictive capability across all classes.
- The **weighted average** F1-score is **0.89**, demonstrating that the model maintains consistency across different obesity categories.

2. Confusion Matrix Analysis

The **confusion matrix** helps visualize the classifier's predictions compared to actual labels.

- **Class 4 (Obesity Type II)** achieved a perfect precision, recall, and F1-score of **1.00**, indicating that the model correctly classified all instances.
- **Class 3 (Obesity Type I) and Class 0 (Normal Weight)** also showed excellent predictive performance, with recall values above **95%**.
- **Class 1 (Overweight Level 1) and Class 5 (Obesity Type III)** had slightly lower F1-scores of **0.79** and **0.76**, suggesting some misclassifications.
- **Class 6 (Underweight) had an F1-score of 0.85**, indicating moderate classification performance for this category.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

3. Accuracy and Performance Evaluation

- The model achieved an overall **accuracy of 89.13%**, reflecting its high effectiveness in obesity classification.
- The **recall values** indicate that the model is capable of detecting most cases correctly, with slight underperformance in some classes.
- The **precision scores** demonstrate that the model effectively avoids false positives in most cases.

4. Feature Importance Analysis

Although SVM does not inherently provide feature importance, **permutation importance** was used to analyze which features contributed the most to predictions. The key features influencing obesity classification were:

1. **BMI (Body Mass Index)** – Strongest predictor.
2. **Caloric Intake** – High impact on overweight and obesity levels.
3. **Physical Activity Levels** – Key determinant for underweight and normal weight classification.
4. **Genetic Factors (Family History of Obesity)** – Influential for higher obesity classes.

5. Visualization of Results

- **Confusion Matrix Heatmap**: A graphical representation of true vs. predicted labels was generated to analyze misclassification trends.
- **Feature Importance Graph**: A bar chart was used to display the most significant predictors contributing to obesity classification.

The SVM model demonstrated **high accuracy (89.13%)**, with strong precision and recall across most obesity categories. The results indicate that **machine learning can effectively classify obesity levels** based on demographic and behavioral factors. Future enhancements could include **ensemble learning techniques (e.g., Random Forest, XGBoost) or deep learning models** to further improve classification performance.

V. CONCLUSION

This study explored the effectiveness of **Support Vector Machine (SVM)** for obesity prediction using the **Kaggle Obesity Prediction Dataset**. The model achieved a high accuracy of **89.13%**, demonstrating its strong predictive capability in classifying individuals into different obesity categories. Performance metrics such as **precision, recall, and F1-score** indicated that the classifier performed well across most classes, particularly for **Obesity Type II**, which was predicted with perfect accuracy. However, slight misclassifications were observed in **Overweight Level 1 and Obesity Type III**, suggesting areas for improvement.

The **confusion matrix heatmap** provided insights into the classification performance, highlighting trends in misclassification. Additionally, **feature importance analysis** revealed that **BMI, caloric intake, and physical activity levels** were the most influential factors in obesity prediction. The findings suggest that **machine learning techniques can effectively identify obesity risks**, aiding in early intervention and personalized health recommendations.

Future work could explore **ensemble models like Random Forest and XGBoost**, or **deep learning architectures** to enhance predictive accuracy. Integrating additional factors such as **genetic predisposition and metabolic data** could further refine classification performance. The research highlights the **potential of AI in healthcare analytics**, offering valuable tools for obesity prevention and management. Implementing these models in **real-world healthcare applications** could significantly improve early detection and treatment strategies.

REFERENCES

[1] K. Joshi, N. Aher, S. Arora, V. Mulay, R. Suryawanshi, N. Pise, and V. Gutte, "Comparison of Different Machine Learning and Self-Learning Methods for Predicting Obesity on Generalized and Gender-Segregated Data," Int. J. Recent Innov. Trends Comput. Commun., vol. 11, no. 10, pp. 464–471, 2023. [Online]. Available: ijritcc.org

[2] Y.-C. Lee, J. J. Christensen, L. D. Parnell, C. E. Smith, J. Shao, N. M. McKeown, J. M. Ordovás, and C.-Q. Lai, "Using Machine Learning to Predict Obesity Based on Genome-Wide and Epigenome-Wide Gene–Gene and Gene–Diet Interactions," Front. Genet., vol. 12, 2022. [Online]. Available: frontiersin.org

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.423

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753; p-ISSN: 2347-6710|

Volume 13, Issue 7, July 2024

|DOI: 10.15680/IJIRSET.2024.1307185|

[3] M. Gupta, T. T. Phan, H. T. Bunnell, and R. Beheshti, "Obesity Prediction with EHR Data: A Deep Learning Approach with Interpretable Elements," ACM Trans. Comput. Healthc., vol. 3, no. 3, p. 32, 2022.
[4] E. Rodríguez, E. Rodríguez, L. Nascimento, A. da Silva, and F. Marins, "Machine Learning Techniques to Predict Overweight or Obesity," Int. J. Data Min. Model. Manag., vol. 13, no. 2, pp. 123–138, 2021.
[5] P. K. Mondal, K. H. Foysal, B. A. Norman, and L. S. Gittner, "Predicting Childhood Obesity Based on Single and Multiple Well-Child Visit Data Using Machine Learning Classifiers," Sensors, vol. 23, no. 3, p. 759, 2023.
[6] S. A. Thamrin, D. S. Arsyad, H. Kuswanto, A. Lawi, and S. Nasir, "Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018," Front. Nutr., vol. 8, p. 669155, 2021.
[7] F. H. Yagin, M. Gülü, Y. Gormez, A. Castañeda-Babarro, C. Colak, G. Greco, F. Fischetti, and S. Cataldi, "Estimation of Obesity Levels with a Trained Neural Network Approach Optimized by the Bayesian Technique," Appl. Sci., vol. 13, no. 7, p. 3875, 2023.
[8] F. Ferdowsy, K. S. A. Rahi, M. I. Jabiullah, and M. T. Habib, "A Machine Learning Approach for Obesity Risk Prediction," Curr. Res. Behav. Sci., vol. 2, p. 100053, 2021.
[9] M. Anisat, S. Patil, and S. K. Jain, "Development of Obesity Prediction Model Using Machine Learning Techniques," Int. J. Sci. Eng. Res., vol. 13, no. 8, pp. 240–245, 2022.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details