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

Machine Learning enabled Obesity Classificaion using Health and Lifestyle Indicators

Khushi Hayaran, Dr. Spandan Ranpariya
Department of Physics, IISHLS, Indus University, Ahemedabad, khushihayaran.23.mphy@ishls.indusuni.ac.in
Department of Physics, IISHLS,Indus University, Ahmedabad, spandankumar.ishls@indusuni.ac.in

***ABSTRACT*— Obesity is a health concern that is rapidly becoming a major issue in terms of individual health and work task performance. This study suggests a Support Vector Machine (SVM)-based obesity classification model using the dataset from various health parameters like BMI, age, weight, height, and lifestyle factors. Our model aims to distinguish individuals into several obesity classifications. Performance comparison work has been done by tuning hyper parameters through GridSearchCV and kernel selection optimization, which allows our model to make more predictions that are accurate. We utilized a confusion matrix to visualize the analysis of prediction outcomes and misclassification patterns. The research findings substantiate the efficacy of SVM in predicting obesity trends, and they provide helpful information towards health awareness initiatives, informing the audience about the possible means. The research highlights the significant role of machine learning in preventive care and provides a larger horizon for expanding the broader impacts into larger populations.**

Keywords*— *Obesity, Machine Learning, Support Vector Machine (SVM), Linear Kernel.

I. INTRODUCTION

Nutrition is a behaviour that should be done consciously to protect human health and increase the quality of life, to take the nutrients needed by the body in sufficient quantities and at appropriate times. Adequate and balanced nutrition is one of the basic conditions for an individual to live healthily, develop economically and socially, and increase the level of welfare [1]. Obesity is a significant health issue affecting individuals of all ages worldwide [2,3]. Defined by the World Health Organization (WHO) as a significant determinant of death and disability, obesity stands out as the fourth most common risk factor in terms of non-communicable diseases, following high blood pressure, dietary risks, and tobacco [4,5]. The issue of obesity has transformed into a global health challenge and is progressively escalating in numerous nations, starting from childhood and adolescence [6,7]. According to a study, the number of overweight adults exceeds one billion and body mass increases across all regions of the world when considering the entire population distribution [8].

As shown in Fig. 1, obesity has adverse effects not only on physical health but also on psychosocial and emotional health [9]. As the

prevalence of overweight and obesity continues to rise, there is an increasing need to develop new methods to predict who will benefit from dietary interventions. Artificial Intelligence (AI) has the potential to solve problems using computer systems integrated with large data sources. In the field of nutrition, AI applications can offer benefits such as effective nutrition planning, interpretation, diet analysis, quality nutrition counselling, and in-depth knowledge about the effects of nutrition on health. Considering the literature studies, Machine Learning (ML) helps to study the obesity data [10, 11]. In this study, four different obesity states are determined, which are based on seven features such as Age, Gender, Height, Weight, Physical activity, BMI, etc. For this work, we have utilized the Support Vector Machine (SVM) method to study obesity data. Support Vector Machines (SVMs) are a widely employed and highly effective machine-learning technique for data classification [12]. SVMs are supervised learning models based on statistical learning theory, which involve learning algorithms that analyse the data for classification and regression analysis. This method transforms the initial input space into a higher-dimensional feature space to enhance the classification process. With SVM, limits can be defined for both linear and nonlinear datasets. Support Vector Machines have become widely favoured due to their capability to identify optimal hyper planes that maximize the separation between classes in the feature space [13].

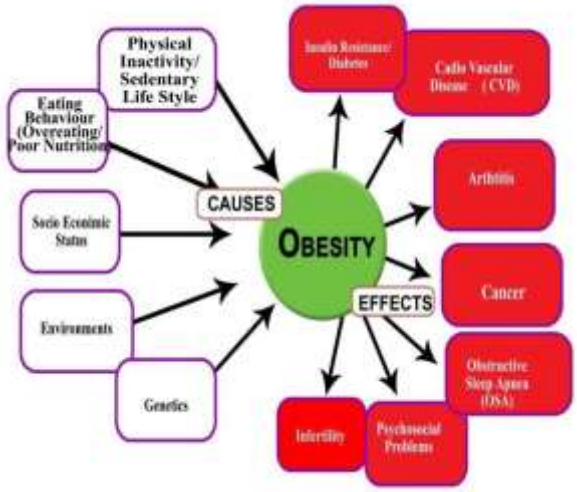

Fig. 1: Various reasons of Obesity [28]

The fundamental principle underlying SVM involves separating classes by drawing margins between them. As illustrated in Fig. 2, these margins are calculated in a manner that maximizes the distance between the margin and the classes, thereby minimizing classification error [14-17].

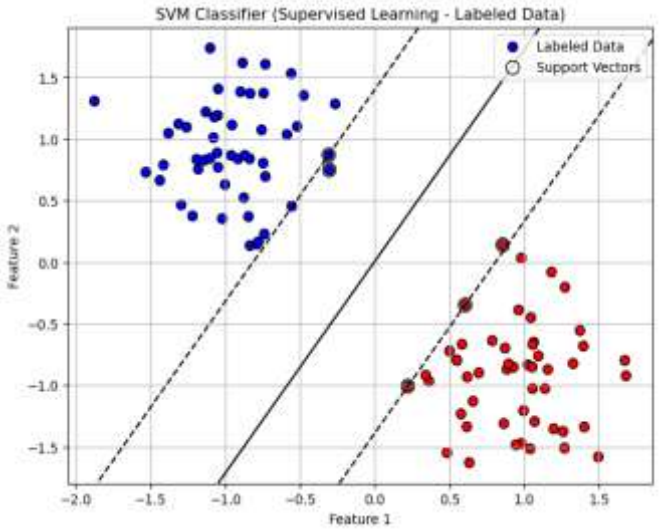

Fig. 2: The decision boundary and cluster representation

Support Vector Machine (SVM) aims to pinpoint the optimal hyper plane that minimizes classification errors while maximizing the margin, a crucial space between data points. This pursuit involves a cost parameter, which balances the desire to maximize accuracy while minimizing misclassifications [18]. Ultimately, SVM strives to delineate a decision boundary that effectively separates different classes of data, ensuring accurate classification of new, unseen data points while guarding against over fitting [19]. To achieve this, SVM employs kernel functions like linear, polynomial, or radial basis functions to transform input data into higher-dimensional spaces. Within these transformed spaces, SVM constructs a hyper plane that efficiently segregates data points into their respective classes, relying solely on a subset of training data points known as support vectors [20].

II. METHODOLOGY

To develop the SVM model, we have utilized the dataset from an open-source website (Kaggle-Obesity Levels Dataset), which includes several physical features such as gender, age, physical activity, height, weight, etc. In the Obesity Category section, the target variable classifies individuals into four categories: underweight, Normal weight, Overweight, and Obese. This dataset was refined by using label encoding, normalization, etc. The development of the algorithm starts with the data pre-processing, in which the data is encoded in categorical features using Label Encoding, through which all string values are converted into numeric values. Further, a dataset was analysed for class imbalance. Since the Kaggle dataset is relatively balanced, no oversampling or under sampling was applied. The whole processed dataset was divided into three parts, 70% training, 15% testing and 15% validation to improve model performance.

In the decision-making process of the linear SVM model, the feature importance was analysed based on the absolute values of the model's coefficients. The bar chart shown in Fig. 3 illustrates that the input variables had the greatest influence on the classification decision. The Support Vector Machine classifier is utilised for its robustness in binary and multiclass classification tasks. The various kernel options, i.e., Linear Kernel, RBF Kernel, Sigmoid Kernel and Polynomial Kernel are studied and the Linear Kernel was chosen for its simplicity and interpretability.

The model's performance was evaluated on the test and validation dataset using the accuracy score and confusion matrix. Additionally, decision boundaries were visualized for the selected features (BMI and Age) to provide an intuitive understanding of model predictions.

To improve computational efficiency and potentially enhance model generalization, Principal Component Analysis (PCA) was applied to reduce the feature space before training the SVM model. PCA helped in projecting the original features onto a lower-dimensional space while retaining maximum variance. The number of components was selected based on explained variance. This step also assisted in visualizing the dataset in a 2D space for explanatory analysis [21].

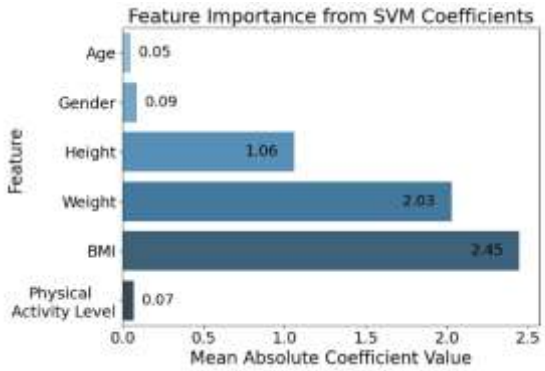

Fig. 3: Importance of features based on average absolute values of the SVM linear model

The model was implemented using the Scikit-learn Python library, proposed by Pedregosa *et al* [22]. A linear support vector machine was used for the classification of obesity levels using labelled data from the obesity dataset. The model was trained by using the soft margin SVM formulation with a regularization parameter C=10, which balances the trade-off between margin maximization and classification.

A soft margin linear SVM can be expressed as:

$$\text{Min } \tfrac{1}{2} \|w\|^2 + C \sum \xi_i$$

Subject to:

$$y_i \ (w^t \ x_i \ + b) \geq 1 - \xi_i \ \geq 0$$

Where, w is weight vector, B is bias term, ξi is slack variables (for misclassification), C is 10 (regularization parameter), $x_i \in$ *Rd input vectors and* $y_i \in$ *{1, 2, 3, 4} is the multiple labels for obesity category.*

The data is linearly separable, so we used a linear kernel.

$$K(x_i, x_j) = x_i^T \ x_j$$

The decision function is also in linear form.

$$f(x) = w^T x + b$$

To manage the four obesity classes, such as underweight, normal, overweight and obese, we used a one-vs-one method, where each classifier distinguishes between a pair of classes. During prediction, a voting mechanism is used to decide the final output class.

III. Results And Discussion

The principal component analysis is utilized to visualise the high-dimensional obesity and understand the natural grouping of different obesity categories. In this, we reduce the dimensionality of data while preserving most of the variance. As shown in Fig. 4, each point represents a data sample projected onto the new coordinate system refined by Principal Component 1 (PC1) and Principal Component 2 (PC2). The samples are color-coded based on their obesity category: Underweight (yellow), Normal weight (purple), Overweight (green), and Obese (blue). It is visible in Fig. 4 that the Underweight individuals are mostly located on the far left, whereas Obese individuals are spread across the right region of the plot, Normal and Overweight category lies in between.

The scatter plot is also plotted which represents the clustering of the dataset based on the output of the trained Linear SVM model as shown in Fig. 5. Each colour denotes a distinct obesity class: Underweight, Normal, Overweight, and Obese. The clear class separation observed in the plot reaffirms that the data is linearly separable, justifying the use of a linear kernel**.**

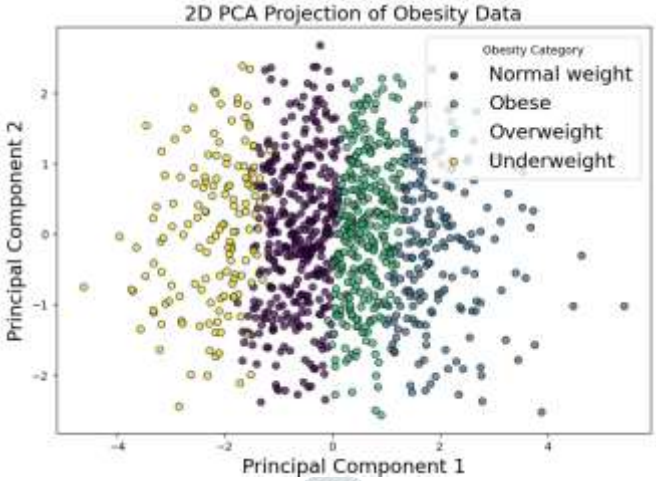

Fig. 4: PCA- based 2D projection of the dataset showing separation by obesity categories.

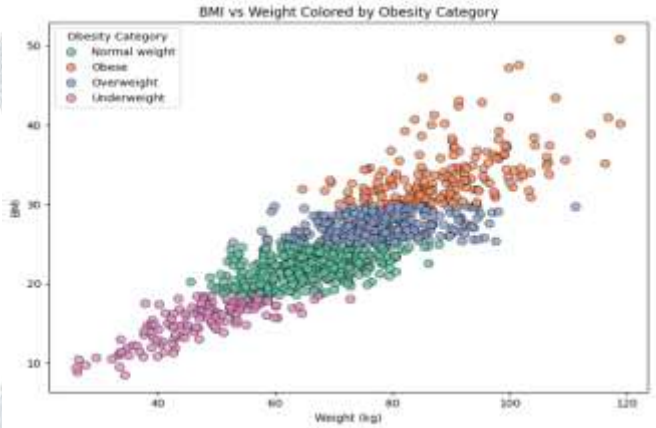

Fig. 5: Scatter plot of BMI versus weight, color-coded by obesity category, represents the features are well separated.

In this study, the multiple SVM kernels are tested, including Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid, to evaluate their performance on the given dataset. All these SVM kernels visualization on Obesity Dataset are shown in Fig. 6. While each kernel was tested under similar conditions to ensure a fair comparison, the linear kernel consistently yielded the highest accuracy and performed best in terms of classification metrics. As a result, the linear kernel was selected for detailed analysis and interpretation throughout this research.

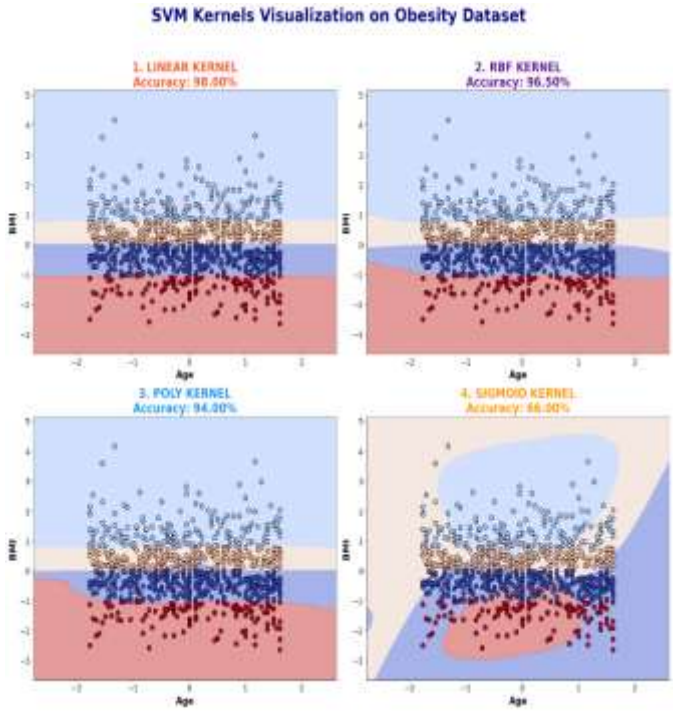

Fig. 6: Different SVM kernels visualization on Obesity Dataset, i.e., 1. Linear Kernel, 2. RBF Kernel, 3. Poly Kernel and 4. Sigmoid Kernel

Support vector machine model is trained with the help of a linear kernel and this linear kernel trained on the obesity dataset achieved an outstanding accuracy. The high performance of a model indicates that the data is likely linearly separable to a large extent, validating the choice of a linear kernel. The regularization parameter C=10 effectively balanced maximum distance between two classes with minimal misclassification. A higher C value restricts the margin to better fit the data points, contributing to this strong accuracy.

Artificial intelligence models can be evaluated using a variety of metrics to assess their performance [23, 24]. The commonly used metrics such as Precision (P), Recall (R), F-measure (F), and Accuracy (AC) were utilized to evaluate the performance of artificial intelligence models. The computation of these metrics for a two-class confusion matrix is as follows; Accuracy: refers to the total number of correctly classified records by a classifier. It is measured as the percentage of correctly classified test sets based on the model, defining how accurate the classifier is [24].

Accuracy (%) = TP + TN/ TP + FP + FN + TN ×100
Where, TP, TN, FP and FN are the True Positive, True Negative, False Positive and False Negative, respectively.

Precision (P) = The proportion of true positive samples to all samples that are classified as positive [14, 24].

Recall(R) = As a measure of accurate identification of positive samples, it refers to the true positive rate [24, 26].
Recall (%) = TP TP + FN ×100

F1-Score (F) = F-measure is calculated by combining precision and recall metrics to evaluate the model's performance [27].

$$F1\text{-}Score = 2 \times TP / 2 \times TP + FP + FN \times 100$$

A four-class confusion matrix was utilized in this study due to the presence of four classes in the dataset. Table 1 depicts the Precision, Recall, F1-Score and Support values of the four-class confusion matrix.

Table 1: The various parameters of the Linear SVM model

Kernel: Linear	**Precision**	**Re-call**	**F1-score**	**support**
Normal weight	0.97	0.97	0.97	72
Obese	1.00	0.97	0.98	32
Overweight	0.95	0.97	0.96	65
Underweight	1.00	1.00	1.00	31

A confusion matrix is a tabular representation implemented for evaluating the accuracy of classification in artificial intelligence methods. The table contains four distinct values, and based on these values, the accuracy of the model can be calculated. The classification using the support vector machine provides an accuracy of ≈ 94.6%. Which shows that the developed model for the dataset can automatically able to predict the obesity in the body with high accuracy.

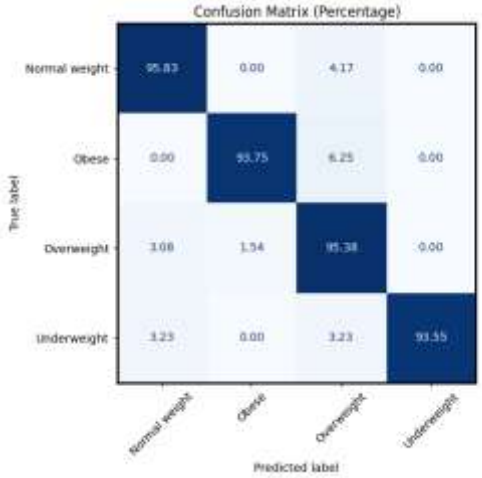

Fig. 7: Confusion matrix output obtained from the developed SVM algorithm to recognize the different types of class, i.e., Normal Weight, Obese, Over Weight and Under Weight.

IV. CONCLUSION

In this paper, Support Vector Machine model is classified the obesity levels based on individual health features such as BMI, weight, height, physical activity etc. This SVM model trained and tested on a dataset of 1001 individuals, enabling reliable generalization across a diverse set of obesity. The visual clustering of classified data provided interpretability and clear separation between the four obesity categories. The different kernel are studied to improve and implement of the best fit type of Kernel. Performance evaluation metrics including precision, recall, F1-score, and a confusion matrix further validated the robustness of the model which confirms that the model has an accuracy of more than 94%. An attempt to implement the machine learning on the obesity dataset has been made and discovered that the machine learning enabled tool can replace the high-cost and end-moment prediction. It is possible to develop a powerful and interpretable tool for obesity classification, particularly in healthcare scenarios where transparency, speed, and accuracy are essential.

ACKNOWLEDGMENT

Khushi Hayaran (K.H.) would like to express the sincere gratitude to her research supervisor, Dr. Spandan Ranpariya, for their constant support, guidance, and encouragement throughout this project. Their insightful feedback played a significant role in shaping the direction of this research.

K.H. would also like to thank Dr. K. C. Roy (Dean, IISHLS, Indus University) and Dr. Manisha Vithalpura (HOD, Physics Department, Indus University) whose unwavering support and motivation helped her to stay committed and focused during this research journey.

Lastly, K.H. is also grateful to her friend Mr. Raj Verma, Mr. Shaurya Hayaran for their backbone support and motivation.

REFERENCES

[1] Singh, P., & Rai, S. N. (2019). Factors affecting obesity and its treatment. Obesity Medicine, 16, 100140. https://doi.org/10.1016/j.obmed.2019.100140

[2] Kopelman, P. G. (2000). Obesity as a medical problem. Nature, 404(6778), 635–643. https://doi.org/10.1038/35007508

[3] Deckelbaum, R. J., & Williams, C. L. (2001). Childhood obesity: the health issue. Obesity Research, 9:239S-243S. https://doi.org/10.1038/oby.2001.125

[4] World Health Organization: WHO. (2024, February 23). The challenge of obesity. (Accessed on 2025, May 15) https://www.who.int/europe/news-room/fact-sheets/item/the-challenge-of-obesity

[5] Vizmanos, B., Cascales, A. I., Rodríguez-Martín, M., Salmerón, D., Morales, E., Aragón-Alonso, A., Scheer, F. A. J. L., & Garaulet, M. (2023). Lifestyle mediators of associations among siestas, obesity, and metabolic health. Obesity, 31(5), 1227–1239. https://doi.org/10.1002/oby.23765

[6] Ogden, C. L., Carroll, M. D., Curtin, L. R., McDowell, M. A., Tabak, C. J., & Flegal, K. M. (2006). Prevalence of overweight and obesity in the United States, 1999–2004. JAMA, 295(13), 1549–1555. https://doi.org/10.1001/jama.295.13.1549

[7] Ng, M., Fleming, T., Robinson, M., Thomson, B., Graetz, N., Margono, C., Mullany, E. C., Biryukov, S., Abbafati, C., Abera, S. F., Abraham, J. P., Abu-Rmeileh, N. M. E., Achoki, T., AlBuhairan, F. S., Alemu, Z. A., Alfonso, R., Ali, M. K., Ali, R., Guzman, N. A., Gakidou, E. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. The Lancet, 384(9945), 766–781. https://doi.org/10.1016/s0140-6736(14)60460-8

[8] Finucane, M. M., Stevens, G. A., Cowan, M. J., Danaei, G., Lin, J. K., Paciorek, C. J., Singh, G. M., Gutierrez, H. R., Lu, Y., Bahalim, A. N., Farzadfar, F., Riley, L. M., & Ezzati, M. (2011). National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and epidemiological studies with 960 country-years and 9.1 million participants. The Lancet, 377(9765), 557–567. https://doi.org/10.1016/s0140-6736(10)62037-5

[9] Reinehr, T. (2010). Obesity and thyroid function. Molecular and Cellular Endocrinology, 316(2), 165–171. https://doi.org/10.1016/j.mce.2009.06.005

[10] Uribe, A. L. M., & Patterson, J. (2023). Are nutrition professionals ready for artificial intelligence? Journal of Nutrition Education and Behavior, 55(9), 623. https://doi.org/10.1016/j.jneb.2023.07.007

[11] Chatterjee, A., Gerdes, M. W., & Martinez, S. G. (2020). Identification of Risk Factors Associated with Obesity and Overweight—A Machine Learning Overview. Sensors, 20(9), 2734. https://doi.org/10.3390/s20092734

[12] Ozkan, I. A., Koklu, M., & Sert, I. U. (2018). Diagnosis of urinary tract infection based on artificial intelligence methods. Computer Methods and Programs in Biomedicine, 166, 51–59. https://doi.org/10.1016/j.cmpb.2018.10.007

[13] Schölkopf, B., & Smola, A. J. (2001). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. https://doi.org/10.7551/mitpress/4175.001.0001

[14] Vapnik, V. N. (2000). The nature of statistical learning theory. In Springer eBooks. https://doi.org/10.1007/978-1-4757-3264-1

[15] Roman, I., Santana, R., Mendiburu, A., & Lozano, J. A. (2020). In-depth analysis of SVM kernel learning and its components. Neural Computing and Applications, 33(12), 6575–6594. https://doi.org/10.1007/s00521-020-05419-z

[16] Valkenborg, D., Rousseau, A., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. American Journal of Orthodontics and Dentofacial Orthopedics, 164(5), 754–757. https://doi.org/10.1016/j.ajodo.2023.08.003

[17] Mahesh, B. (2020). Machine learning algorithms—a review. International Journal of Science and Research (IJSR), 9(1), 381–386. https://doi.org/10.21275/ART20203995

[18] Wang, Q. (2022). Support Vector machine algorithm in machine learning. 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 750–756. https://doi.org/10.1109/icaica54878.2022.9844516

[19] Zhu, F., Zhang, W., Chen, X., Gao, X., & Ye, N. (2023). Large margin distribution multi-class supervised novelty detection. Expert Systems With Applications, 224, 119937. https://doi.org/10.1016/j.eswa.2023.119937

[20] Patle, A., & Chouhan, D. S. (2013, January 23–25). SVM kernel functions for classification. In Proceedings of the International Conference on Advances in Technology and Engineering (ICATE 2013) (pp. 202–206). Mumbai, India. https://doi.org/10.1109/ICAdTE.2013.6524743

[21] Yu, H., & Kim, S. (2012). SVM tutorial—classification, regression and ranking. In G. Rozenberg, T. Bäck, & J. N. Kok (Eds.), Handbook of Natural Computing (pp. 479–506). Springer. https://doi.org/10.1007/978-3-540-92910-9_15

[22] Meriga, B., Ganjayi, M. S., & Parim, B. N. (2017). Phytocompounds as potential agents to treat obesity-cardiovascular ailments. Cardiovascular & Hematological Agents in Medicinal Chemistry, 15(2), 104–120. https://pubmed.ncbi.nlm.nih.gov/28875833

[23] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011, October 12). SciKit-Learn: Machine Learning in Python. https://hal.science/hal-00650905/

[24] Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2018). Deep recurrent neural network for intrusion detection in SDN-based networks. In 2018 IEEE International Conference on Network Softwarization (NetSoft) (pp. 202–206). https://doi.org/10.1109/NETSOFT.2018.8460090

[25] Unal, Y., Taspinar, Y. S., Cinar, I., Kursun, R., & Koklu, M. (2022). Application of Pre-Trained Deep Convolutional Neural Networks for coffee beans species detection. Food Analytical Methods, 15(12), 3232–3243. https://doi.org/10.1007/s12161-022-02362-8

[26] Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). In Springer series in statistics. https://doi.org/10.1007/b98835

[27] Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29, 685–693. https://doi.org/10.1007/s00521-016-2604-1