Performance Analysis of Machine Learning-Based Models for Early Diagnosis of Obesity Using Blood Test Parameters

Sare Nur Cuhadar*, Gul Karaduman**‡, Ahmet Uyanık***, Habibe Durmaz*

* Karamanoğlu Mehmetbey University, Department of Electrical and Electronics Engineering, Karaman, TÜRKİYE, 70200

** Karamanoğlu Mehmetbey University, Vocational School of Health Services, Karaman, TÜRKİYE, 70200

*** Konya Meram State Hospital, Konya, TÜRKİYE, 42020

(sarenurcuhadar22@gmail.com, gulk@bu.edu, uyanik.md@gmail.com, durmazgul@gmail.com)

‡ Corresponding Author; Gul KARADUMAN, Karamanoğlu Mehmetbey University, Vocational School of Health Services, 70200 Karaman, TÜRKİYE, Tel: +90 338 226 2761,

Fax: +90 338 226 2190, gulk@bu.edu

Received: 24.10.2023 Accepted:27.11.2023

Abstract- Obesity is a global health issue that continues to grow, with projections indicating further increases in obesity rates. The World Health Organization defines overweight and obesity as the abnormal or excessive accumulation of fat, posing risks to overall health. Obesity is not only a significant condition itself but is also directly linked to various diseases such as type 2 diabetes, coronary heart disease, hypertension, and certain types of cancer. The rising prevalence of obesity presents significant health complications and risks for individuals of all ages, particularly children and adolescents. Obese or overweight children face an increased likelihood of developing severe health problems in adulthood, potentially enduring the same physical condition throughout their lives. Urgent action is necessary to mitigate this global health concern. In this study, we aimed to predict obesity risk through the use of machine learning algorithms. Our research gathered 367 data from individuals of different age groups, classified as either obese or non-obese, based on their blood test results. We employed nine machine learning algorithms, including BayesNet, Naïve Bayes, SMO, Simple Logistic, IBk, Kstar, J48, Random Forest, and Random Tree algorithms. Our analysis successfully determined the obesity status of individuals based on internal results, the Simple Logistic algorithm achieved the highest accuracy rate at 98.6395. On the other hand, the Simple Logistic and Kstar algorithm demonstrated the highest accuracy rate of 100% for the extenal set. Our model provides valuable insights for further research and interventions for analyzing the blood test values associated with obesity.

Keywords Obesity, machine learning, feature selection, classification, model performance.

1. Introduction

The global obesity crisis is intensifying, with a projected significant increase in its prevalence rates both within individual countries and across the world in the coming years. According to the World Health Organization (WHO), overweight and obesity are described as the abnormal or excessive accumulation of fat that has the potential to have adverse effects on one's health. The global obesity epidemic has emerged as a critical and rapidly growing health concern, especially in developed countries. Obesity is a major risk factor for a variety of chronic illnesses, and it can significantly reduce one's quality of life [1,2]. It is not only a serious condition itself, but it is also closely linked to a multitude of health conditions, including type 2 diabetes, coronary heart disease, hypertension, and several forms of cancer. Thus, it has emerged as one of the most pervasive public health challenges in the contemporary world, carrying substantial implications for both individual well-being and public healthcare systems. Obesity and overweight have been recognized as global epidemics, affecting both developed and developing countries both for adults and children [3]. In 2020, 39 million children under the age of 5 were identified as either overweight or obese.

In this article, we focus on the classification of obesity using Data Mining that can accurately predict obesity by using selected/meaningful parameters from patients [4]. The primary purpose of the method we used is to identify important data in decision making systems after going through several methods and processes. Data mining is widely used across various fields, benefiting all disciplines [5]. Data distribution among classes within datasets is termed a classification algorithm. These algorithms are trained on distribution patterns from provided training sets. Subsequently, test data is integrated for classification, and the algorithm accurately applies the learned process. Labels determine data set classes in training and test groups, ensuring systematic classification [6].

In 2014, Borrell and Samuel investigated the influence of overweight and obesity on mortality rates among adults in the United States. They utilized a regression method that considered highly correlated hazards to estimate the progression rates for all causes. Their analysis of mortality rates revealed that overweight and obese individuals faced a CVD-related mortality rate that was more than 20% higher than that of normal-weight individuals [9]. In a study conducted in 2015 by Dugan et al., the diagnosis of obesity in children aged two and older was assessed using six different testing models: Random Tree, Random Forest, ID3, J48, Naïve Bayes, and Bayes Net. These models were trained using CHICA, a clinical decision support system. Among these models, the ID3 model demonstrated superior performance, achieving an impressive 85% accuracy and approximately 90% precision [10]. In 2018, Jindal et al. employed machine learning techniques to predict obesity levels with a remarkable accuracy of 89.68%. They harnessed the Python interface and employed generalized linear models, random forest, and partial least squares methods as part of their prediction models [11]. In 2020, Singh and Tawfik introduced a machine learning approach for forecasting the likelihood of obesity or being overweight in adolescents. They utilized seven machine learning algorithms, including k-NN, J48 pruned tree, Random forest, Bagging, support vector machine, multilayer perceptron, and voting. Among these, the MLP algorithm exhibited a sensitivity of 96%, with an impressive F1 score of 93.96% [12]. In 2021, Faria Ferdowsy and colleagues aimed to predict the risk of obesity using machine learning algorithms. They explored nine different algorithms, including k-nearest neighbor (k-NN), random forest, logistic regression, multilayer perceptron (MLP), support vector machine (SVM), naïve Bayes, adaptive boosting (ADA boosting), decision tree, and gradient boosting classifier algorithms. The experiments covered high, moderate, and low obesity classes. Notably, the logistic regression algorithm achieved the highest accuracy rate at 97.09%, while the gradient-boosting classifier algorithm recorded the lowest accuracy, at 64.08% [13]. In 2019, Xueqin Pang and colleagues utilized the XGBoost model, a machine learning approach, to predict the likelihood of early childhood obesity at the age of 2. The model was applied to both males and females, resulting in an AUC of 81% (±0.1%). Risk factors associated with obesity were further examined through the interpretation of XGBoost model predictions. The models were tailored for various age ranges when obesity incidence typically occurs. For individuals within the 24-36 month range, the model achieved a substantial accuracy of 97.63%, while for the 72-84 month age range, the obesity incidence prediction stood at 48.96% .

The objective of our study was to predict the risk of obesity using machine learning algorithms. 367 data was collected from diverse age groups, including both obese and non-obese individuals. The dataset was constructed based on blood test results obtained from the study participants. We employed nine machine learning algorithms, namely BayesNet, Naïve Bayes, SMO, Simple Logistic, IBk, Kstar, J48, Random Forest, and Random Tree algorithms to predict obesity risk. The dataset was split into training and test data for evaluation purposes to determine the obesity status of individuals accurately[14].

2. Material and Methods

In this study, we employed WEKA software to develop models for obesity prediction [15]. WEKA encompasses various tools for tasks like data preprocessing, association rule mining, classification, clustering, prediction, visualization, and regression [16]. Fig. 1. outlines the steps in creating machine learning models to diagnose obesity from blood test parameters.

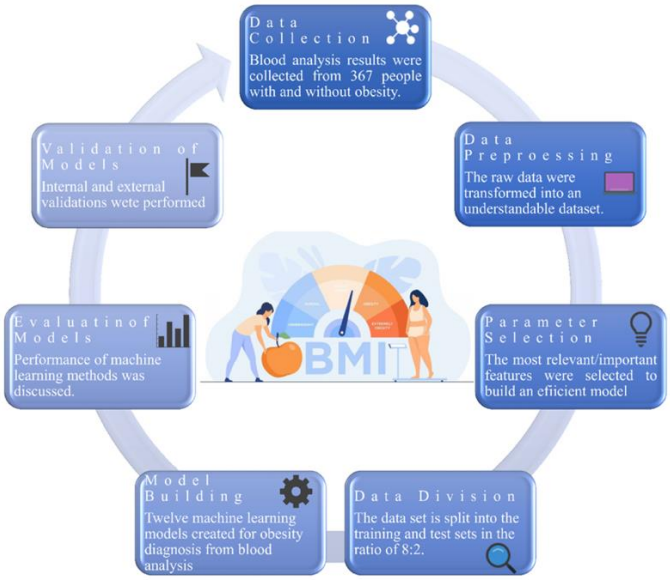

Fig. 1. Machine learning model development

We began with dataset preprocessing, entailing noise removal, normalization, and applicability domain determination. Subsequently, we divided the dataset into a training set and a test set at an 8:2 ratio. To pinpoint the optimal parameters for accurate obesity prediction from blood test results, we deployed nine machine learning algorithms, evaluating them using confusion matrices for internal and external validation.

2.1. Dataset

Table 1 offers our dataset, featuring 367 instances with 23 attributes. This dataset was instrumental in our study employing the WEKA program for obesity diagnosis and prediction. It encompasses blood test results for obesity diagnosis, categorized as "obesity" and "non-obesity" based on disease-associated values.

Detailed information on the descriptive values of the dataset used in the study is explained below.

- ID: The dataset was created using the numbering system of the blood results.
- *Height:* The height of the individuals.
- *Weight:* The weight of the patients.
- *Age:* The age of the individuals.
- *Gender:* This attribute includes the gender of the. Males are represented in the dataset as 0, and females are represented as 1.
- *Iron:* It is a type of mineral that enables red blood cells to carry oxygen to the body and is mostly found in hemoglobin and myoglobin pigments [17].
- *TIC:* Total iron binding capacity (TIBC) is a medical laboratory test that measures how well iron-binding sites in serum can be saturated [18].
- *LDL Cholesterol:* Low-density lipoprotein (LDL) is commonly referred to as bad cholesterol [19].
- *Cholesterol:* Cholesterol is a waxy lipid, a type of fat found in human blood and produced naturally by all cells, especially in the liver [20].
- *HDL Cholesterol:* High-density lipoprotein (HDL) is commonly referred to as good cholesterol [21].
- *Triglycerides:* Triglycerides are another type of lipid, different from cholesterol [22].
- *ALT:* Alanine Aminotransferase (ALT) is an enzyme predominantly found in the liver, which the body uses to convert food into energy [23].
- *AST:* Aspartate Aminotransferase is an enzyme produced by the liver, and the AST test is a biochemical laboratory test used to detect liver damage and diagnose many liver diseases [23].
- *Glucose:* Glucose is the body's main source of energy, and excess is stored as glycogen in the liver and muscles.
- *Creatinine:* Creatinine is an amino acid produced naturally in the body and found in very small amounts in food [24].
- *CRP:* C-Reactive protein (CRP) is a type of protein produced by the liver in response to inflammation anywhere in the body [21].
- *CK:* Creatine kinase (CK) is the main enzyme responsible for the production of the creatine phosphate molecule [25].
- *HbA1c:* The HbA1c test result reflects the average blood glucose level over the last 2 to 3 months. Specifically, the HbA1c test measures the percentage of hemoglobin, the oxygen-carrying protein of red blood cells, that is coated with sugar [26].
- *Vitamin B12:* Vitamin B12 is a water-soluble B vitamin that is needed by almost every cell in the body due to its role in DNA synthesis and the utilization of fatty acids and amino acids in the body [27].
- *TSH:* Thyroid stimulating hormone is secreted by the pituitary gland. It enters the thyroid gland through the bloodstream and stimulates the gland [28].
- *Ferritin:* Ferritin is a protein complex responsible for storing and utilizing iron minerals obtained from nutrients [29].
- *Urea:* Urea is a molecule that removes nitrogenous substances from the body and is cleared from the blood through the kidneys.
- *Uric Acid:* Uric acid is an organic compound composed of carbon, oxygen, nitrogen, and hydrogen. It is formed by the oxidation of oxypurinol by xanthine oxidase [30].

Table 1. Characteristics of the obesity diagnostic dataset

Data Source	Feature Abbreviation	Feature Description	Feature Reference Range	Feature Unit
Patient Numbering	ID	Patient Numbering	-	-
Patient Physical Characteristics	Size	The height of the patients	141-193	cm
	Weight	Body weight of the patients	33-190	kg
Patient Age	Age	Patient life expectancy	11-86	J

Patient Gender	Gender	Women:1 Men:0	-	-
Patient blood results	Iron	Mineral	50 – 170	ug/dl
	TIBC	Total Iron-Binding Capacity	70 - 130	ug/dl
	LDL Cholesterol	Low-density lipoprotein	0.0 - 130	mg/dl
	Cholesterol	Lipoprotein	125 - 250	mg/dl
	HDL Cholesterol	High-density lipoprotein	0 - 60	mg/dl
	Triglyceride	Lipid	50 - 150	mg/dl
	ALT	Alanine aminotransferase	4 - 44	U/L
	AST	Aspartat transaminaz	7 – 38	U/L
	Glucose	Fasting Blood Glucose	70 – 100	mg/mL
	Creatinine	Amino Acid	0.6 – 1.3	mg/dl
	CRP	C Reactive Protein	0.0 – 5	mg/L
	CK	Creatine Kinase	29 – 168	U/L
	HbA1c	Glycolyzed hemoglobin	3.9 – 6.1	%
	Vitamin B12	Vitamin	120 - 883	pg/mL
	TSH	Thyroid Stimulating Hormone	0.35 – 4.94	uIU/ml
	Ferritin	Protein	4.63 – 204	ng/mL
	Urea	Molecule	7 – 18.7	mg/dl
	Uric Acid	Organic Compound	2.6 – 6.0	mg/dl

2.2. Data Preprocessing

In this stage of the study, we created a structured data table from the collected data, ensuring the completeness of the dataset. We excluded any data points with missing values and removed any categorical variables that may cause errors during modeling in WEKA. The preprocessing and data cleaning phase is fundamental for the success of data mining, and its effectiveness directly impacts the quality of the results. We removed rows with a significant number of missing or invalid data points, and for rows with a small number of missing data, we replaced them with the average of the corresponding data group. Additionally, we removed duplicate data instances from the dataset to ensure that only unique data points remained. We also removed irrelevant data such as name and surname to enhance the quality of the results. Finally, we applied normalization by rescaling the data between 0 and 1 to improve modeling accuracy [31].

2.3. Data Division

Our dataset was split into two segments for training and testing. 80% of the data served as the training set, enabling the algorithm to learn data relationships. Subsequently, we used the remaining 20% as the test set to evaluate the algorithm's performance. This allowed us to assess the algorithm's accuracy independently from the training data [32].

2.4. Attribute Selection

Attribute selection enables the identification of the most significant attributes in a dataset during the development of a new model [33]. It also allows for the selection of features that are correlated with one another. In our study, we utilized the information and correlation-based feature evaluators (CfsSubsetEvals) provided by the WEKA program to identify values associated with obesity.

2.5. Machine Learning Algorithms

2.5.1. Naive Bayes

The Naive Bayes classification algorithm is a probabilistic approach based on Bayes' theorem. Its peculiarity is that it assumes independence between variable values, which may not reflect reality. This assumption, known as class conditional independence, helps define classification parameters when working with small training datasets [34]. The classification algorithm combines decision rules with Bayesian probability models, hence the term 'naive'.

2.5.2. Random Forest

The Random Forest classification algorithm creates multiple classification trees using subsets of the dataset and then combines them to improve classification accuracy. Each tree is constructed using a randomly sampled vector of features from the input data [35]. To construct the decision forest, individual decision trees are combined. In the Random Forest algorithm, new training sets are formed by replacing the original datasets, and each tree is created using a randomly selected subset of attributes. Nodes are split using the maximum split value from the randomly selected attributes [36].

2.5.3. Decision Tree (J48)

The decision tree (J48) classification algorithm is a straightforward approach to data classification using decision trees. At each node, the algorithm selects the attribute that partitions the data into classes most efficiently. This method does not require the discretization of numerical attributes and can handle missing values and continuous attribute value ranges. Additionally, the J48 algorithm has features such as rule derivation, pruning decision trees, and the ability to work with both categorical and continuous dependent variables [37].

2.5.4. Support Vector Machines

The Support Vector Machines (SVM) classification algorithm is a pairwise approach that employs decision boundaries to classify a given class. It is a machine learning method based on a vector field that identifies the decision boundary between the two classes farthest from a random point in the training data. SVM is generally used to classify datasets with two classes. However, for datasets with multiple classes, the basic SVM algorithm may not be adequate. The main objective of SVM is to identify a hyperplane in n-dimensional feature space that can uniquely classify data points. It identifies the area with the maximum margin, which is the maximum distance between the data points in both classes. It helps to maximize the margin range [38].

2.5.5. K-nearest neighbors classifier (K-NN)

The k-nearest neighbors classification algorithm (k-NN) is a non-parametric algorithm commonly used in supervised learning for classification and regression tasks. It is relatively straightforward to implement and does not have many assumptions about the underlying data. The algorithm calculates the minimum distance point of the nearest neighbor and uses distance functions to calculate the distances of the new data points to the existing dataset. Finally, the algorithm assigns labels to the data points based on the nearest neighbors [39].

2.5.6. Logistic Regression

Logistic regression is a commonly used classification algorithm, particularly in binary classification tasks. The algorithm summarizes the linear and additive effect of variable properties on the probability of an event using regression coefficients [40]. Logistic regression models can have one or more independent variables that determine the outcome. The algorithm returns two possible results, 1 (true) or 0 (false). It is commonly used to analyze data and describe the relationship between a dependent binary variable and one or more independent variables.

2.5.7. Multilayer Perceptron

The multilayer perceptron algorithm is considered the starting point of the deep learning method. It is mostly used in artificial neural networks, which consist of a minimum of two or more layers of nodes [41].

2.5.8. Random Tree

The Random Tree algorithm is a fast decision tree learner and a supervised classifier. To build a decision tree, the algorithm generates a random dataset using the idea of bagging. The tree is then pruned using the reduced error pruning method with backpropagation [42].

2.6. Evaluation of the Models

There are many evaluation metrics used to measure the performance of a machine learning model, depending on the specific task and type of model. The metrics we use are based on our problem and the goals of the model and include accuracy (ACC), Precision, Recall, F-Measure, Matthews correlation coefficient (MCC), receiver operating characteristics (ROC), and Kappa statistics. All the parameters were calculated based on the confusion matrix, which consists of the true positive (TP), true negative (TN), false positive (FP), and false negative (FN) numbers.

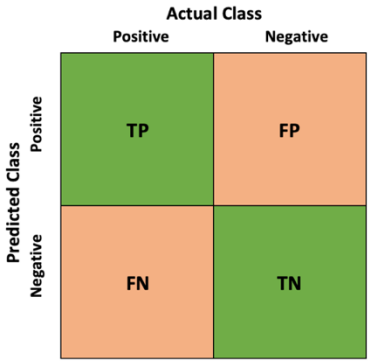

Fig. 2. Confusion matrix for binary classification

True Positive: It is called the true positive rate. It is when the disease is found, and the diagnostic test gives a positive result.

True Negative: It is the true negative rate. It is when the diagnostic test result of the individual who is not sick is also not sick.

False Positive: False positive rate, where the disease is not present. But the diagnostic test gives a positive result.

False Negative: False negative rate is when the diagnostic test of an individual who is ill gives the result that the individual is not ill.

Accuracy: It is defined as the percentage of correct predictions made by a diagnostic system when new data is added to it [43].

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}. \quad (1)$$

Precision: Indicates how many of the values predicted with a positive result in a generated system are actually positive [44].

$$Precision = \frac{TP}{(TP + FP)}. \quad (2)$$

Recall: It is the parameter that shows how many of the values that should be predicted positively in the system are predicted positively [45].

$$Recall = \frac{TP}{(TP + FN)}. \quad (3)$$

F-Measure: The F-measure is defined as the Recall Average and Harmonic Sensitivity.

It can be expressed by F, and its equation is as follows [46],

$$F - Measure = 2 * \left(\frac{Precision * Recall}{Precision + Recall}\right). \quad (4)$$

Receiver Operating Characteristic Curve: The curve that represents the performance of classifiers is called the ROC curve. ROC curves are plotted to evaluate the test samples with a high positive rate. They provide a visual representation of the classifier's performance, irrespective of the type of error or class distribution. The curve plots the number of positive samples on the horizontal axis (x-axis) as a percentage of the number of negative samples, while the vertical axis shows the percentage of negative samples corresponding to the number of positive samples. ROC categories: Excellent-[0.90-1], Good-[0.80-0.90], Fair-[0.70-0.80], Bad-[0.60-0.70] and Poor-[0.50-0.60] [47].

Matthews Correlation Coefficient (MCC): MCC is used to evaluate the effectiveness of binary categorizations by analyzing the imbalance of positive and negative cases. The perfect MCC value for prediction is 1.0, while the worst MCC value is 0.0. The formula for calculating MCC is as follows [48].

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \quad (5)$$

Kappa statistics: They are statistical properties that examine the interaction between the ratios of two categorized data sets. The value of the datasets can range from 0 to 1, or even higher in the case of stronger ratios. If the value of K is equal to 1, this indicates that the agreement between classifiers is within a tolerable range and should be accepted as an accurate value. If the value of K is equal to 0, then there is no agreement. Kappa statistics are represented by the symbol K is formulated as follows.

$$K = \frac{P(A) - P(E)}{1 - P(E)}, \quad (6)$$

where K = Kappa Statistics , P (A) = Percentage of agreement , P (E) = Chance of agreement [49].

3. Results

3.1. Feature Selection

Several experiments were conducted on the dataset we created using a variety of nine machine-learning techniques. Through feature selection the most important biomarkers associated with obesity disease were identified based on blood values. Once the biomarkers were identified, the dataset was split into training and test sets. Table 2 presents a list of features selected from the dataset, which were considered during the experiments conducted in our study. These selected features include height, weight, age, cholesterol, alanine aminotransferase (ALT), and aspartate transaminase (AST).

Table 2. The selected parameters

Number of Parameters	Parameters
6	height
	weight
	age
	cholestrol
	Alanine aminotransferase (ALT)
	Aspartate transaminase (AST)

3.2. Internal Validation

Table 3 presents the results of the training data in the dataset, which consisted of 294 samples. Analysis of the classification results shows that the Simple Logistic and J48 classifiers achieved higher correct classification rates (98.6395% - 290 correct classifications and 96.5986% - 284 correct classifications, respectively) compared to the other classifiers employed. Moreover, the Simple Logistic classifier had a significantly lower misclassification rate (1.3605% and 4 misclassifications) in comparison to the IBk classifier (9.8639% and 29 misclassifications).

Table 3. The correctly and incorrectly classified instances and the accuracy for the internal validation set

Classifier	NCCI	NICI	ACC
BayesNet	269	25	91.4969 %
Naïve Bayes	270	24	91.8367 %
SMO	268	26	91.1565 %
Simple Logistic	290	4	98.6395 %
IBk	265	29	90.1361 %
Kstar	273	21	92.8571 %
J48	284	10	96.5986 %
Random Forest	283	11	96.2585 %
Random Tree	282	12	95.9184 %

NCCI: Number of correctly classified instances, NICI: Number of incorrectly classified instances, ACC: Accuracy

According to Table 4, Simple Logistic classification has the highest TP Rate (0.986), while IBk has the lowest TP Rate (0.901). Additionally, Simple Logistic and J48 classifiers exhibit the best Precision values (0.987 and 0.966, respectively). It is worth noting that the BayesNet classification has the highest FP Rate (0.243).

Table 4. TP Rate, FP Rate, Precision, Recall rates of the internal validation set

Classifier	TP Rate	FP Rate	Precision	Recall
BayesNet	0.915	0.243	0.913	0.915
Naïve Bayes	0.918	0.126	0.922	0.918
SMO	0.912	0.232	0.909	0.912
Simple Logistic	0.986	0.015	0.987	0.986
IBk	0.901	0.189	0.902	0.901
Kstar	0.929	0.170	0.927	0.929
J48	0.966	0.067	0.966	0.966
Random Forest	0.963	0.114	0.963	0.963
Random Tree	0.959	0.115	0.959	0.959

TP Rate: True Positive Rate, FR Rate: False Positive Rate

Table 5 indicates that Simple Logistic and J48 classifiers have the highest MCC values (0.960 and 0.899, respectively), while IBk has the lowest MCC value (0.734). Moreover, the Simple Logistic classifier has the highest percentage of the ROC area, which is 99.3%.

Table 5. F-Measurement, MCC, ROC Area, and Kappa Statistic

Classifier	F-measure	MCC	ROC	Kappa Statistic
BayesNet	0.911	0.734	0.925	0.727
Naïve Bayes	0.920	0.767	0.934	0.7657
SMO	0.912	0.725	0.840	0.7213
Simple Logistic	0.986	0.960	0.993	0.9601
IBk	0.902	0.709	0.932	0.7088
Kstar	0.927	0.783	0.967	0.7816
J48	0.966	0.899	0.964	0.899
Random Forest	0.962	0.887	0.990	0.8842

Random Tree	0.958	0.876	0.922	0.8744

MCC: Matthew's Correlation Coefficient, ROC: Receiver Operating Characteristic Curve

3.3. External Validation

Table 6 displays the results of the data extracted and tested from the dataset. A total of 73 samples were utilized for classification testing. The outcomes of the test reveal that Simple Logistic, Kstar, and Random Forest classifiers exhibit higher correct classification rates (100%-73%, 100%-73%, and 98.6301%-72%, respectively) compared to the other classifiers. Furthermore, the misclassification rate of the Simple Logistic classifier (0% and 0 misclassifications) is lower in contrast to the BayesNet and J48 classifiers (6.8493% and 5 misclassifications).

Table 6. The correctly and incorrectly classified instances and the accuracy of the external validation set

Classifier	NCCI	NICI	ACC
BayesNet	68	5	93.1507 %
Naïve Bayes	69	4	94.5205 %
SMO	70	3	95.8904 %
Simple Logistic	73	0	100 %
IBk	71	2	97.2603 %
Kstar	73	0	100 %
J48	68	5	93.1507 %
Random Forest	72	1	98.6301 %
Random Tree	69	4	94.5205 %

NCCI: Number of correctly classified instances, NICI: Number of incorrectly classified instances, ACC: Accuracy

Table 7 displays that Simple Logistic and Kstar classifications exhibit the highest TP Rate (1.000), whereas BayesNet and J48 exhibit the lowest TP Rate (0.932). Additionally, Simple Logistic and Kstar classifications have the best Precision (1.000 and 1.000, respectively). The BayesNet classification displays the highest FP Rate (0.234).

Table 7. TP Rate, FP Rate, Precision, and Recall rates of the external validation set

Classifier	TP Rate	FP Rate	Precision	Recall
BayesNet	0.932	0.234	0.930	0.932
Naïve Bayes	0.945	0.122	0.945	0.945
SMO	0.959	0.173	0.961	0.959
Simple Logistic	1.000	0.000	1.000	1.000
IBk	0.973	0.007	0.976	0.973
Kstar	1.000	0.000	1.000	1.000
J48	0.932	0.125	0.934	0.932
Random Forest	0.986	0.003	0.987	0.986
Random Tree	0.945	0.176	0.944	0.945

TP Rate: True Positive Rate, FR Rate: False Positive Rate

Table 8 shows that Simple Logistic and Kstar have the highest MCC values (1.000 each), while BayesNet has the lowest MCC value (0.767). Additionally, Simple Logistic and Kstar have the highest percentage of the ROC area (100%).

Table 8. F-Measurement, MCC, ROC Area, and Kappa Statistic

Classifier	F-measure	MCC	ROC	Kappa Statistic
BayesNet	0.928	0.767	0.927	0.7594
Naïve Bayes	0.945	0.823	0.967	0.8232
SMO	0.957	0.865	0.893	0.8556
Simple Logistic	1.000	1.000	1.000	1
IBk	0.973	0.919	0.987	0.9162
Kstar	1.000	1.000	1.000	1
J48	0.932	0.786	0.959	0.7849
Random Forest	0.986	0.958	0.999	0.957
Random Tree	0.944	0.817	0.884	0.8131

MCC: Matthew's Correlation Coefficient, ROC: Receiver Operating Characteristic Curve

3.4. Correlation Analysis of the Selected Prarameters

In Fig. 3., we established a statistical relationship between parameters for the early diagnosis of obesity. We plotted a correlation heatmap using Python version 3.11 (https://www.python.org/) to visualize the pairwise correlations. The heatmap shows that Pearson's correlation

values can range from -1 to 1, and negative correlations are indicated with brown, while positive correlations are indicated with dark green.

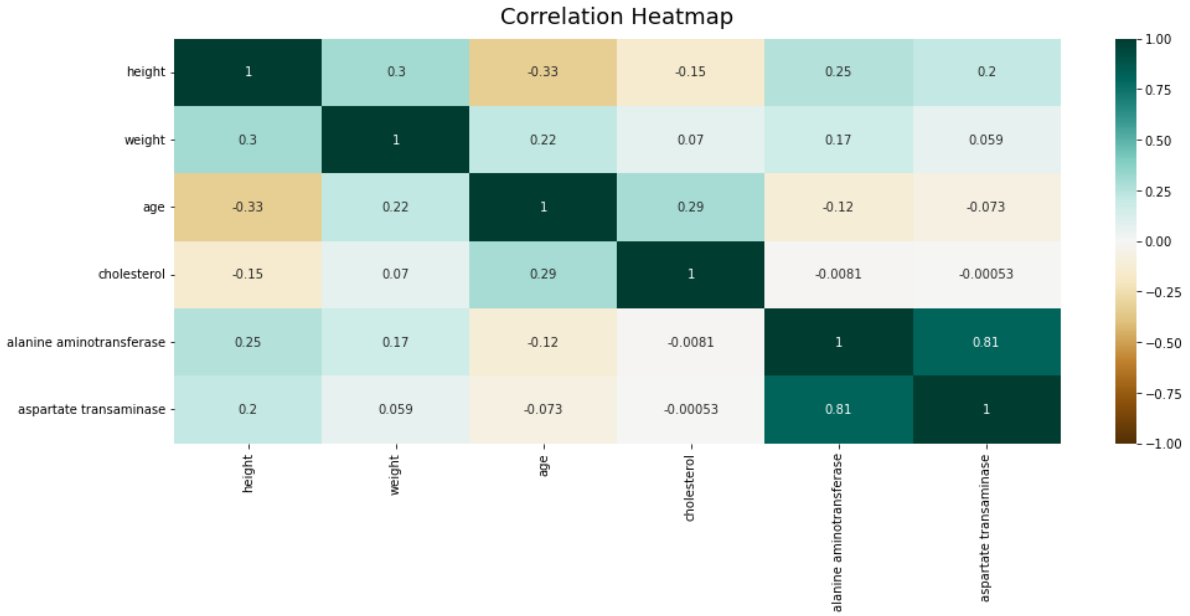

Fig. 3. Correlation analysis of the selected parameters for early diagnosis of obesity

Based on the heatmap, we found that parameter pairs with strong positive correlations include aspartate transaminase & alanine aminotransferase. Parameter pairs with positive correlations include weight & height, height & alanine aminotransferase, aspartate transaminase & height, age & weight, cholesterol & age, and alanine aminotransferase & weight. However, we did not find any parameter pair with strong negative correlations. We also noticed that some parameters have low correlations or no correlations, including weight & cholesterol, aspartate transaminase & weight, alanine aminotransferase & cholesterol, and aspartate transaminase & cholesterol. Finally, we found that some parameters have negative correlations but are not strongly related, including age & height, cholesterol & height, alanine aminotransferase & age, and age & aspartate transaminase.

4. Discussion

The validation of machine learning models is crucial to assess their performance and reliability in predicting obesity risk. In our study, we conducted several experiments using a dataset created from individuals classified as obese or non-obese, focusing on biomarkers associated with obesity disease based on blood values. The selected features for experimentation included height, weight, age, cholesterol, alanine aminotransferase (ALT), and aspartate transaminase (AST).

(ROC) area, indicating its superior discriminative ability (99.3%). These results demonstrate that the Simple Logistic algorithm performs exceptionally well in accurately predicting obesity risk.

During the internal validation, which involved the training data consisting of 294 samples, the Simple Logistic and J48 classifiers achieved the highest correct classification rates (98.6395% and 96.5986%, respectively) compared to the other classifiers. The Simple Logistic classifier also had a significantly lower misclassification rate (1.3605%) compared to the IBk classifier (9.8639%). These results indicate that the Simple Logistic and J48 algorithms are more effective in accurately classifying individuals as obese or non-obese based on the selected features. Furthermore, the True Positive (TP) Rate and Precision values were analyzed to evaluate the performance of the classifiers. The Simple Logistic classifier demonstrated the highest TP Rate (0.986) and Precision (0.987), indicating its ability to identify individuals who are truly obese correctly. On the other hand, the IBk classifier had the lowest TP Rate (0.901) and Precision (0.902). These findings suggest that the Simple Logistic classifier is more reliable in correctly classifying obese individuals. Matthews's Correlation Coefficient (MCC) was also examined to assess the overall performance of the classifiers. The Simple Logistic and J48 classifiers exhibited the highest MCC values (0.960 and 0.899, respectively), while the IBk classifier had the lowest MCC value (0.734). The Simple Logistic classifier also had the highest percentage of the Receiver Operating Characteristic

For external validation, using a separate dataset consisting of 73 samples, the Simple Logistic, Kstar, and Random Forest classifiers achieved higher correct classification rates (100%, 100%, and 98.6301%, respectively) compared to the other classifiers. The Simple

Logistic classifier had a perfect accuracy rate with zero misclassifications, demonstrating its robustness in classifying individuals correctly. Conversely, the BayesNet and J48 classifiers had higher misclassification rates (6.8493% and 5 misclassifications), indicating a comparatively lower performance. The TP Rate and Precision values for external validation revealed that the Simple Logistic and Kstar classifiers achieved the highest TP Rate (1.000) and Precision (1.000). In contrast, the BayesNet classifier had the lowest TP Rate (0.932), while the Simple Logistic and Kstar classifiers had the highest MCC values (1.000 each). These findings further emphasize the superior performance of the Simple Logistic classifier in correctly identifying individuals at risk of obesity.

The study highlighted that Simple Logistic and J48 classifiers exhibited strong performance in internal validation, achieving high correct classification rates. Simple Logistic, in particular, stood out with the highest TP Rate, Precision, MCC, and ROC area percentage, emphasizing its accuracy. External validation further supported these findings, where Simple Logistic, Kstar, and Random Forest achieved superior correct classification rates, and Simple Logistic achieved perfect accuracy. Overall, the research underscored the effectiveness and reliability of the Simple Logistic algorithm for predicting obesity risk, suggesting its potential use in clinical settings for early detection and prevention of obesity-related complications.

5. Conclusion

In conclusion, our study employed machine learning algorithms to predict the risk of obesity using blood test results. The findings demonstrate the potential of machine learning techniques in accurately classifying individuals as obese or non-obese based on blood values. Early identification of individuals at risk of obesity can facilitate targeted interventions and preventive measures to address this global health issue effectively. Further research and validation are warranted to enhance the accuracy and applicability of these predictive models in real-world settings.

Ethics statement

This retrospective observational study was conducted in compliance with the principles of the Declaration of Helsinki and was approved by the ethics committee of Karamanoglu Mehmetbey University School of Medicine. Due to its retrospective nature, the requirement for informed consent was waived by the ethics board of Karamanoglu Mehmetbey University School of Medicine.

Conflict of interest

The authors declare that there is no conflict of interest regarding the content of this paper.

References

[1] K. Nimptsch and P. Tobias. “Body fatness, related biomarkers and cancer risk: an epidemiological perspective.” Hormone molecular biology and clinical investigation 22, no. 2, pp. 39-51, 2015.

[2] Prospective Studies Collaboration. “Body-mass index and cause-specific mortality in 900 000 adults: collaborative analyses of 57 prospective studies.” The Lancet 373, no. 9669, pp. 1083-1096, 2009.

[3] L. Cominato, G.F. Di Biagio, D. Lellis, R.R. Franco, M.C. Mancini, and M.E. de Melo. “Obesity prevention: strategies and challenges in Latin America.” Current obesity reports 7, pp. 97-104, 2018.

[4] J.H. Friedman. “Data Mining and Statistics: What's the connection?.” Computing science and statistics, 29(1), pp. 3-9, 1998.

[5] F.E. Horita, J.P. de Albuquerque, V. Marchezini, and E.M. Mendiondo. “Bridging the gap between decision-making and emerging big data sources: An application of a model-based framework to disaster management in Brazil.” Decision Support Systems, 97, pp. 12-22, 2017.

[6] S.B. Kotsiantis, I. Zaharakis, and P. Pintelas. “Supervised machine learning: A review of classification techniques.” Emerging artificial intelligence applications in computer engineering, 160(1), pp. 3-24, 2007.

[7] J. Sun and C.K. Reddy. “Big data analytics for healthcare.” In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1525-1525, 2013..

[8] K. Srinivas, B.K Rani, and A. Govrdhan. “Applications of data mining techniques in healthcare and prediction of heart attacks.” International Journal on Computer Science and Engineering (IJCSE), 2(02), pp. 250-255, 2010.

[9] L.N. Borrell and L. Samuel. “Body mass index categories and mortality risk in us adults: the effect of overweight and obesity on advancing death.” Am. J. Public Health 104 (3), 2014.

[10] T.M. Dugan, S. Mukhopadhyay, A. Carroll, S. Downs. “Machine learning techniques for prediction of early childhood obesity.” Appl. Clin. Inform. 6 (3), 2015.

[11] K. Jindal, N. Baliyan, and P.S. Rana. “Obesity prediction using ensemble machine learning approaches.” In: Proceedings of the 5th ICACNI, 2, pp. 355–362, 2017.

[12] B. Singh and H. Tawfik. “Machine learning approach for the early prediction of the risk of overweight and obesity in young people.” In Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20, pp. 523-535, Springer International Publishing, 2020.

[13] F. Ferdowsy, K.S.A Rahi, M.I. Jabiullah, and M.T. Habib. “A machine learning approach for obesity risk prediction.” Current Research in Behavioral Sciences, 2, 100053, 2021.

[14] A.M. Erturan, G. Karaduman, and H. Durmaz. “Machine learning-based approach for efficient prediction of toxicity of chemical gases using feature selection.” Journal of hazardous materials, 455, 131616, 2023.

[15] E. Frank, M.A. Hall, and I.H. Witten. “The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques".” 4th edn. Morgan Kaufmann, Burlington, 2016.

[16] S.R. Garner. “Weka: The waikato environment for knowledge analysis.” In Proceedings of the New Zealand computer science research students conference, pp. 57-64, 1995.

[17] M.Ç .Cengiz. “Obezite cerrahi geçiren bireylerde yağ dokusu kaybı ile demir ve D vitamini düzeyi arasındaki ilişki.” Master's thesis, Biruni Üniversitesi Sağlık Bilimleri Enstitüsü, 2019.

[18] K.G. Şişman and Ş.K. Anemisi. “Beslenme Örüntüsü ile Kronik İnflamasyon Belirteçleri ve Diyet Tedavisinin Etkinliğinin Belirlenmesi.” Doktora tezi. Ankara: Hacettepe Üniversitesi, 2013.

[19] I. Damoune, I. Khaldouni, L. Agerd and F. Ajdi. “Obésité: prévalence et profil métabolique chez une population de diabétique type 2.” In Annales d'Endocrinologie, Vol. 75, No. 5-6, pp. 457, Elsevier Masson, 2014.

[20] M. Valle, R. Martos, F. Gascon, R. Canete, M.A. Zafra and R. Morales. “Low-grade systemic inflammation, hypoadiponectinemia and a high concentration of leptin are present in very young obese children, and correlate with metabolic syndrome.” Diabetes & metabolism, 31(1), pp. 55-62, 2005.

[21] M. Valle, R. Martos, F. Gascon, R. Canete, M.A. Zafra, and R. Morales. “Low-grade systemic inflammation, hypoadiponectinemia and a high concentration of leptin are present in very young obese children, and correlate with metabolic syndrome.” Diabetes & metabolism, 31(1), pp. 55-62, 2005.

[22] H. Hüsna. “Santral obezite ve bel/kalça çevresinin dislipidemi ile ilişkisi.” Dünya Beslenme Dergisi , 1 (2), pp. 18-22, 2018.

[23] M.A. Burza, S. Romeo, A. Kotronen, P.A. Svensson, K. Sjöholm, J.S. Torgerson, and M. Peltonen. “Long-term effect of bariatric surgery on liver enzymes in the Swedish Obese Subjects (SOS) study.” PloS one, 8(3), e60495, 2013.

[24] J. A. Demirovic, A.B. Pai, and M.P. Pai. “Estimation of creatinine clearance in morbidly obese patients.” American Journal of Health-System Pharmacy, 66(7), pp. 642-648, 2009.

[25] J. J Rayner, M.A. Peterzan, W.D. Watson, W.T. Clarke, S. Neubauer, C.T. Rodgers, and O.J. Rider. “Myocardial energetics in obesity: enhanced ATP delivery through creatine kinase with blunted stress response.” Circulation, 141(14), pp. 1152-1163, 2020.

[26] B. Hansel, P.Giral, L.Gambotti, A.Lafourcade, G. Peres, C.Filipecki, D.Kadouch, A.Hartemann, J.M. Oppert, E. Bruckert, and M. Marre. “A fully automated web-based program improves lifestyle habits and HbA1c in patients with type 2 diabetes and abdominal obesity: randomized trial of patient e-coaching nutritional support (the ANODE study).” Journal of medical Internet research, 19(11), pp.e360, 2017.

[27] O. Pinhas-Hamiel, N. Doron-Panush, B. Reichman, D. Nitzan-Kaluski, S. Shalitin, and L. Geva-Lerner. “Obese children and adolescents: a risk group for low vitamin B12 concentration.” Archives of pediatrics & adolescent medicine, 1;160(9), pp. 933-6, 2006.

[28] A. Valea, M. Carsote, C. Moldovan, and C. Georgescu. “Chronic autoimmune thyroiditis and obesity.” Archives of the Balkan Medical Union, 53(1), pp. 64-69, 2018.

[29] A. SÜNER, O. BALAKAN, V. KIDIR. “Association of Thalassemia Minor and Lead Intoxication in a Patient who Applied with Hypochromic Microcytic Anemia.” International Journal of Hematology and Oncology, 32(1), pp. 133-136, 2006.

[30] N.H. Noğay and G. Köksal. “Çocuklarda metabolik sendromun tedavisinde beslenme yönetimi” Güncel Pediatri, 10(3), pp. 92-97, 2012.

[31] F. Kelleci Çelik and G. Karaduman. “In silico QSAR modeling to predict the safe use of antibiotics during pregnancy.” Drug and Chemical Toxicology. doi: 10.1080/01480545.2022.2113888, pp. 1-10, 2002.

[32] G. Karaduman and F. Kelleci Çeli. “2D-Quantitative structure-activity relationship modeling for risk assessment of pharmacotherapy applied during

pregnancy." Journal of Applied Toxicology: JAT, 10.1002/jat.4475. https://doi.org/10.1002/jat.4475, 2023.

[33] F. Kelleci Çelik and G. Karaduman. "Machine Learning-Based Prediction of Drug-Induced Hepatotoxicity: An OvA-QSTR Approach." Journal of Chemical Information and Modeling, 63(15), pp. 4602-4614, 2023.

[34] M. Narasimha Murty, V. Susheela Devi, "Pattern Recognition: An Algorithmic Approach", Springer Science & Business Media, May 25, 2011.

[35] K. Sridharan and G. Komarasamy. "Sentiment classification using harmony random forest and harmony gradient boosting machine." Soft Computing, 24(10), pp. 7451-7458, 2020.

[36] Z. Wang, F. Chegdani, N. Yalamarti, B. Takabi, B. Tai, M. El Mansori, and S. Bukkapatnam. "Acoustic Emission Characterization of Natural Fiber Reinforced Plastic Composite Machining Using a Random Forest Machine Learning Model." Journal of Manufacturing Science and Engineering, 142(3), 2020.

[37] N. Bhargava, G. Sharma, R. Bhargava, and M. Mathuria. "Decision tree analysis on j48 algorithm for data mining." Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 3(6), 2013.

[38] G. Fung and O. L. Mangasarian. "Incremental support vector machine classification." In Proceedings of the 2002 SIAM International Conference on Data Mining pp. 247- 260, Society for Industrial and Applied Mathematics, 2002.

[39] S. Li, K. Zhang, Q. Chen, S. Wang, and S. Zhang. "Feature Selection for High Dimensional Data Using Weighted K-Nearest Neighbors and Genetic Algorithm." IEEE Access, 2020.

[40] F. C. Pampel. "Logistic Regression: A Primer", SAGE Publishers, pp. 35-39, May 26, 2000.

[41] P. Perner. "Machine Learning and Data Mining in Pattern Recognition", 6th International Conference, MLDM 2009, Leipzig, Germany, July 23-25, 2009.

[42] A. Cutler and G. Zhao. "Pert-perfect random tree ensembles." Computing Science and Statistics, 33, pp. 490-497, 2001.

[43] GBD 2019 Mental Disorders Collaborators. Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Psychiatry, 9(2), pp. 137-150, 2022.

[44] R.S.C. Aman. "Disease predictive models for healthcare by using data mining techniques: state of the art." SSRG Int. J. Eng. Trends Technol, 68, pp. 52-57, 2020.

[45] A. Muniasamy, V. Muniasamy, and R. Bhatnagar, "Predictive analytics for cardiovascular disease diagnosis using machine learning techniques," in Advances in Intelligent Systems and Computing, vol. 114, pp. 493–502, 2021.

[46] J. Majali, R. Niranjan, V. Phatak, O. Tadakhe. "Data Mining Techniques for Diagnosis And Prognosis of Cancer", Int. Journal of Advanced Research in Computer and Communication Engg., Vol. 4, Issue 3, pp. 613-614, 2015.

[47] A.P. Sinhaand and J.H. May. "Evaluating and tuning predictive data mining models using receiver operating characteristic curves." Journal of Management Information Systems, 21(3), pp. 249-280, 2005.

[48] K. R. Lakshmi, M. Veera Krishna, S.Prem Kumar. "Performance Comparison of Data Mining Techniques for Prediction and Diagnosis of Breast Cancer Disease Survivability", Asian Journal of Computer Science and Information Technology, Vol. 3, pp. 81 – 87, 2013.

[49] P. Apostolou and F. Fostira. "Hereditary Breast Cancer: The Era of New Susceptibility Genes", BioMed Research International Vols. 2013.