Biointerface Research in Applied Chemistry
Open-Access Journal (ISSN: 2069-5837)

Article
Volume 15, Issue 3, 2025, 39
https://doi.org/10.33263/BRIAC153.039

Integrating QSAR Analysis and Machine Learning to Explore the Antidiabetic Potential of Natural Compounds

Bahar Sincar [1], Dilek Yalcin [2], Oguz Bayraktar [1,*]

[1] Department of Bioengineering, Ege University, Bornova-İzmir, Turkey; baharsncr11@gmail.com (B.S.); oguz.bayraktar@ege.edu.tr (O.B.);
[2] Department of Bioengineering, Izmir Institute of Technology, Urla-İzmir, Turkey; dilekyalcin84@gmail.com (D.Y.);
* Correspondence: oguz.bayraktar@ege.edu.tr;

Received: 10.02.2025; Accepted: 10.05.2025; Published: 8.06.2025

Abstract: This study explores the antidiabetic potential of 72 natural compounds using molecular descriptors and QSAR modeling combined with machine learning techniques. The dataset includes 11 experimentally obtained compounds and 61 from the literature, characterized by their IC_{50} values indicating 50% inhibition of α-glucosidase enzyme activity. Molecular descriptors were generated using ChemAxon's MarvinSketch and PADEL software, narrowing down over 3000 descriptors to 23 relevant features. Statistical analysis revealed significant multicollinearity among variables, necessitating the application of non-linear machine learning models, namely Random Forest and Gradient Boosting. These models demonstrated predictive capabilities with R^2 values of 0.7751 and 0.8066, respectively, and highlighted molecular weight and the number of heteroatoms in ring structures as critical features influencing IC_{50} values. Despite the dataset's variability and limited size, the study underscores the potential of integrating QSAR and machine learning approaches to effectively predict the antidiabetic activity of natural compounds. The findings provide valuable insights for advancing computational methods in drug discovery.

Keywords: antidiabetic potential; QSAR modeling; machine learning; natural compounds; α-glucosidase inhibition.

1. Introduction

Diabetes mellitus (DM) is a chronic metabolic disorder affecting millions worldwide and represents a significant public health concern. As of 2021, the International Diabetes Federation (IDF) reported that approximately 537 million adults globally were living with diabetes, a figure projected to reach 783 million by 2045 [1]. This rising prevalence highlights an urgent need for effective therapeutic agents to manage diabetes and its associated complications. Despite the availability of pharmacological treatments, their long-term use is often linked to adverse effects such as gastrointestinal discomfort, lactic acidosis, and drug resistance [2]. Furthermore, the economic burden of diabetes treatment, particularly in low- and middle-income countries, highlights the need to explore alternative and cost-effective therapeutic approaches [3].

One promising avenue in diabetes management is using plant-derived polyphenolic compounds, which have gained attention due to their structural diversity, bioavailability, and

relatively low toxicity. Polyphenols, including flavonoids and phenolic acids, have demonstrated antidiabetic properties through multiple mechanisms, such as enhancing insulin secretion, inhibiting carbohydrate-digesting enzymes, reducing glucose absorption, and exerting antioxidant and anti-inflammatory effects [4,5].

Flavonoids, a major subclass of polyphenols, include compounds such as quercetin, rutin, epicatechin, and catechin. These bioactive molecules are found in various fruits, vegetables, tea, and wine. Quercetin, for example, has been reported to improve insulin sensitivity and reduce oxidative stress, while rutin has shown potential in lowering blood glucose levels and modulating lipid metabolism [6,7]. Epicatechin, a predominant polyphenol in green tea and cocoa, has been linked to improved glucose uptake and insulin secretion [8,9]. Catechin, another abundant flavonoid, has demonstrated antioxidant and anti-inflammatory properties that contribute to better glycemic control [9,10].

Phenolic acids, another important category of polyphenols, include compounds such as caffeic acid, vanillic acid, rosmarinic acid, and trans-cinnamic acid. These molecules are widely distributed in coffee, herbs, and spices. Caffeic acid has been associated with reduced glucose absorption and enhanced insulin action [11,12]. Vanillic acid, present in vanilla and almonds, has been shown to modulate oxidative stress and inflammation, both of which are implicated in diabetes progression [13-15]. Rosmarinic acid, found in rosemary and thyme, enhances insulin signaling pathways and provides neuroprotective benefits against diabetes-associated complications [16,17]. Trans-cinnamic acid, a compound found in cinnamon and garlic, has demonstrated glucose-lowering properties by modulating key metabolic pathways [18,19].

Despite their potential, the effectiveness of polyphenolic compounds in diabetes treatment requires further systematic evaluation. Traditional drug discovery approaches rely heavily on experimental screening, which is resource-intensive and time-consuming. Computational approaches, such as Quantitative Structure-Activity Relationship (QSAR) modeling, have emerged as valuable tools in predicting the biological activities of chemical compounds based on molecular descriptors [20]. QSAR models establish the relationship between molecular properties and biological activity, enabling researchers to identify promising drug candidates efficiently [21].

Traditional QSAR methods, such as Multiple Linear Regression (MLR) and Partial Least Squares (PLS), have been used to establish predictive models. However, these methods often struggle with complex datasets and non-linear interactions inherent in biological systems. Recent advancements in machine learning (ML) algorithms, including Random Forest (RF) and Gradient Boosting (GB), have demonstrated superior predictive power by effectively handling non-linearity and multicollinearity in molecular data [22,23]. RF models, in particular, have outperformed conventional QSAR techniques in identifying key molecular features governing biological activity, while GB algorithms have shown robust accuracy, particularly in small datasets [24,25].

In parallel, integrated approaches combining *in vitro* assays with computational modeling have gained attention for their ability to elucidate molecular mechanisms and support QSAR predictions. For instance, Salau *et al*. investigated the inhibitory effects of betulinic acid on diabetes-related digestive enzymes using both experimental and *in silico* methods, highlighting the synergistic value of dual validation strategies in antidiabetic drug discovery [26].

This study investigated the antidiabetic potential of 72 natural polyphenolic compounds with reported α-glucosidase inhibitory activities (IC_{50} values). The dataset included 11 experimentally obtained compounds and 61 compounds extracted from the literature, comprehensively representing structurally diverse molecules. Molecular descriptors were generated using ChemAxon's MarvinSketch and PADEL software, yielding over 3000 features, which were subsequently reduced to 23 key descriptors based on their relevance and interpretability [27].

Statistical analysis revealed significant multicollinearity among molecular descriptors, necessitating the application of advanced machine-learning models for accurate predictions. RF and GB algorithms were implemented, achieving R^2 values of 0.7751 and 0.8066, respectively, underscoring their effectiveness in capturing complex relationships between molecular descriptors and IC_{50} values. Notably, molecular weight (MW) and the number of heteroatoms in ring structures (nHeteroRing) emerged as the most influential factors governing α-glucosidase inhibition [27,28].

Moreover, feature importance analysis provided mechanistic insights into molecular interactions with α-glucosidase, revealing that higher molecular weight compounds with specific heteroatom configurations exhibited stronger inhibitory effects. This suggests potential optimization strategies for the design of more potent antidiabetic agents [29].

By integrating QSAR modeling with machine learning techniques, this study contributes to the rapid identification and evaluation of naturally derived antidiabetic compounds. The findings provide a foundation for further *in vitro* and *in vivo* investigations, ultimately facilitating the development of more effective diabetes treatments [30]. Such computational approaches accelerate drug discovery and enhance our understanding of the structure-activity relationships governing bioactive molecules in diabetes therapeutics.

This study aims to investigate the antidiabetic potential of 72 natural polyphenolic compounds by integrating Quantitative Structure-Activity Relationship (QSAR) modeling with machine learning algorithms. By correlating molecular descriptors with α-glucosidase inhibitory activity (IC_{50} values), the study seeks to identify key structural features influencing bioactivity and to develop predictive models that can guide the discovery of novel, natural antidiabetic agents.

2. Materials and Methods

2.1. Selection of polyphenolic compounds and acquisition of chemical descriptors for quantitative structure-property relationships (QSPR).

The selected polyphenols are as follows: morin (1), vanillic acid (2), curcumin (3), rutin (4), epicatechin (5), coumarin (6), catechin (7), caffeic acid (8), trans-cinnamic acid (9), p-coumaric acid (10), and rosmarinic acid (11). The standards of these polyphenols (with a purity of 96% or higher) were used in the experiments. Additionally, the molecular structures of these polyphenols are provided in Figure 1.

The molecular structures of the selected compounds were drawn using the MarvinSketch online tool (ChemAxon) with SMILES (Simplified Molecular Input Line Entry Specification) data to represent them in short ASCII sequences and saved as files in .mol format. These files were analyzed using open-source software such as PADEL. The PADEL software generated more than 2,700 descriptor data for each compound and prepared for use in QSAR (Quantitative Structure-Activity Relationship) analyses. These data facilitated the

identification of structural features associated with the biological activities of polyphenols and enabled the development of related predictive models.

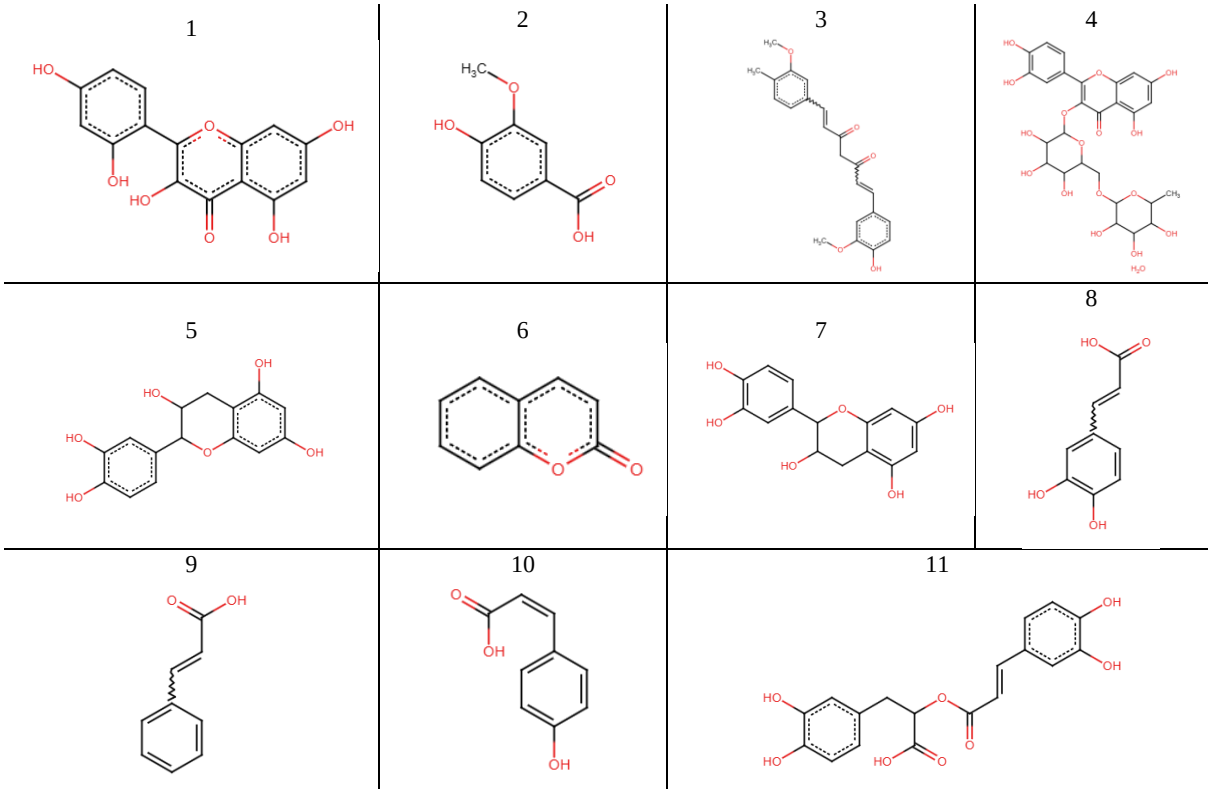


Figure 1. Structures of selected polyphenols for this study.

2.2. In vitro α-glucosidase enzyme ınhibition test.

The α-glucosidase enzyme inhibition test, which has been widely utilized in various graduate-level studies conducted at the natural bioactive materials laboratory, was also performed in this study following the same established protocols.

The test began by dissolving the selected polyphenol standards in 2% DMSO. The α-glucosidase enzyme (1 U/ml) and 5 mM 4-Nitrophenyl-alpha-D-glucopyranoside substrate were prepared in 0.1 M phosphate buffer (pH 6.9). Acarbose and the test samples were added to a 96-well plate in a volume of 50 µl in triplicate. Subsequently, 10 µl of the enzyme was added to each sample and control, followed by the addition of 50 µl of 0.1 M phosphate buffer. The plate was incubated in the dark at 37°C, shaking at 30 rpm for 15 minutes.

After pre-incubation, 20 µl of the substrate was added to each well. The plate was then incubated under the same conditions (37°C, 30 rpm, dark) for 10 minutes. The reaction was terminated by adding 50 µl of 0.1 M sodium carbonate solution to each well. Absorbance measurements were taken at 405 nm using a microplate reader [31-33].

In this study, molecular descriptors and the concentrations at which the α-glucosidase enzyme activity, known to play a role in biochemical processes associated with diabetes, is inhibited by 50% (IC_{50}, µg/ml) were compiled for 72 natural compounds. Of these, 11 were obtained experimentally, and 61 were compiled from studies in the literature. The reference studies from which the IC_{50} (µg/ml) values used as output data in this study were obtained are listed in the References section (these values can also be presented in a table with the references).

2.3. Obtaining molecular descriptors used as ınput data.

Initially, the 2D structures of the 72 natural compounds used in the study were generated using SMILES codes with the MarvinSketch software by ChemAxon (https://marvinjs-demo.chemaxon.com/latest.html). Using the same software, all compounds were saved in the .mol file format.

Subsequently, the .mol file formats were processed using the open-source software PADEL (Pharmaceutical Data Exploration Laboratory), which generated over 3000 molecular descriptive features [27, 34]. These features include the number of atoms, number of bonds, presence of aromatic structures, logP (water solubility) values, characteristic volume, molecular weight, and fingerprint values of the molecules. However, since the majority of these features are not interpretable data, the total of over 3000 molecular descriptors for each compound was narrowed down to 23 selected descriptors. These descriptors are listed in Table 1.

Table 1. Descriptors and their definitions.

No	Descriptor	Definition
1	nAcid	Number of acidic groups
2	naAromAtom	Number of aromatic atoms
3	nAtom	Number of atoms
4	nHeavyAtom	Number of atoms excluding hydrogen
5	nC	Number of carbon atoms
6	nN	Number of nitrogen atoms
7	nO	Number of oxygen atoms
8	nBonds	Number of bonds
9	CrippenLogP	Crippen logP (water solubility)
10	CrippenMR	Crippen molar refractivity
11	nHBAcc	Number of hydrogen bond acceptors
12	nHBDon	Number of hydrogen bond donors
13	HybRatio	Hybridization ratio
14	McGowan_Volume	Characteristic volume
15	nRing	Number of ring structures
16	n5Ring	Number of 5-membered carbon rings
17	nHeteroRing	Number of rings containing atoms other than carbon
18	n5HeteroRing	Number of 5-membered rings containing atoms other than carbon
19	nRotB	Number of rotatable bonds
20	RotBFrac	Fraction of rotatable bonds
21	TopoPSA	Topological polar surface area
22	MW	Molecular weight
23	XLogP	Computed logP (water solubility)

2.4. Data processing and machine learning model development with Python.

The complete dataset used in this study is provided in Table 2. The Python code was utilized for data preprocessing, statistical analyses, model training, and validation tests.

Table 2. Input and output data.

Name	nAcid	naArom Atom	nAto m	nHeavy Atom	nC	nN	nO	nBonds	Crippen LogP	Crippen MR
caffeic acid	1	6	21	13	9	0	4	13	1,25	48,62
catechin	0	12	35	21	15	0	6	23	1,95	76,68
coumarin	0	6	17	11	9	0	2	12	1,81	47,12
Curcumin	0	12	49	27	22	0	5	28	4,12	107,44
Epicatechin	0	12	35	21	15	0	6	23	1,95	76,68
Morin	0	12	32	22	15	0	7	24	1,94	79,84
p-coumaric acid	1	6	20	12	9	0	3	12	1,54	46,96
Rosmarinic acid	1	12	42	26	18	0	8	27	1,81	91,98

Name	nAcid	naAromAtom	nAtom	nHeavyAtom	nC	nN	nO	nBonds	CrippenLogP	CrippenMR
Rutin	0	12	73	43	27	0	16	47	-1,59	141,52
Trans-cinnamic acid	1	6	19	11	9	0	2	11	1,84	45,29
Vanillic acid	1	6	20	12	8	0	4	12	1,07	44,64
(-)-4?-O-Methylepigallocatechin	0	12	39	23	16	0	7	25	1,95	83,24
(-)-epigallocatechin	0	18	51	33	22	0	11	36	1,86	111,16
3-oxolupenal	0	0	78	32	30	0	2	36	7,57	133,01
Apigenin 7-glucuronide	1	12	50	32	21	0	11	35	0,91	110,76
Apigenin	0	12	30	20	15	0	5	22	2,71	74,83
Baicalein	0	12	30	20	15	0	5	22	2,71	74,83
Baicalin	1	12	50	32	21	0	11	35	1,04	110,40
Berberine	0	16	43	25	20	1	4	29	4,15	97,52
Betulinic Acid	1	0	81	33	30	0	3	37	7,25	135,58
Brachystamide B	0	6	67	30	26	1	3	31	6,84	128,36
Chrysin-7-O-glucuronide	1	12	49	31	21	0	10	34	1,20	109,09
Chrysin	0	12	29	19	15	0	4	21	3,00	73,17
Deoxynojirimycin	0	0	24	11	6	1	4	11	-2,97	36,90
Epigallocatechin Gallate	0	18	51	33	22	0	11	36	1,86	111,16
Eriodictyol	0	12	33	21	15	0	6	23	1,84	74,10
Fisetin tetramethyl ether	0	12	43	25	19	0	6	27	3,59	95,45
Galangin	0	12	30	20	15	0	5	22	2,89	74,74
Genistein 5-O-glucuronide	1	12	50	32	21	0	11	35	0,54	108,34
Genistein	0	12	30	20	15	0	5	22	2,71	74,83
Glycycoumarin	0	12	47	27	21	0	6	29	4,52	106,67
Guineensine	0	6	61	28	24	1	3	29	6,06	119,13
Herbacetin	0	12	32	22	15	0	7	24	2,30	78,07
isolicoflavonol	0	12	44	26	20	0	6	28	4,40	101,69
isoliquiritigenin	0	12	31	19	15	0	4	20	2,49	73,20
isoorientin	0	12	52	32	21	0	11	35	0,01	111,13
isoquercitrin	0	12	53	33	21	0	12	36	-0,44	110,29
isovitexin	0	12	51	31	21	0	10	34	0,31	109,46
Kaempferol	0	12	31	21	15	0	6	23	2,59	76,41
Karanjin	0	15	34	22	18	0	4	25	4,31	85,56
Katononic acid	1	0	79	33	30	0	3	37	7,60	134,65
Kotalanol	0	0	50	26	12	0	12	26	-5,70	78,81
Licochalcone A	0	12	47	25	21	0	4	26	4,55	103,46
Liquiritigenin	0	12	31	19	15	0	4	21	2,43	70,77
Liquiritin	0	12	52	30	21	0	9	33	-0,10	103,50
Lupeol	0	0	81	31	30	0	1	35	7,95	130,38
Luteolin	0	12	31	21	15	0	6	23	2,41	76,50
Mangiferin	0	12	48	30	19	0	11	33	-0,09	102,76
Medicarpin	0	12	34	20	16	0	4	23	3,60	76,67
Naringenin	0	12	32	20	15	0	5	22	2,14	72,44
Neoliquiritin	0	12	52	30	21	0	9	33	-0,10	103,50
Orientin	0	12	52	32	21	0	11	35	-0,36	108,71
Ovalitenone	0	15	39	25	19	0	6	28	4,09	93,69
Pellitorine	0	0	41	16	14	1	1	15	3,54	72,97
Phloretin	0	12	34	20	15	0	5	21	2,29	75,22
Pinnatin	0	15	34	22	18	0	4	25	4,34	86,16
Pipataline	0	6	49	21	19	0	2	22	6,28	89,83
Piperonylic acid	1	6	18	12	8	0	4	13	1,58	44,61
Polydatin	0	12	50	28	20	0	8	30	0,53	100,10
Pongachromene	0	12	46	28	22	0	6	32	4,98	107,72
Pongaglabrone	0	15	33	23	18	0	5	27	4,56	87,80
Pongamol	0	15	36	22	18	0	4	24	3,87	85,51
Pongapin	0	15	37	25	19	0	6	29	4,54	93,75
Procyanidin A2	0	24	66	42	30	0	12	48	3,76	148,39
Prunetin-5-O-beta-D-glucuronide	0	12	52	33	22	0	11	36	-0,49	110,59
Prunetin	0	12	33	21	16	0	5	23	3,01	79,72
Salacinol	0	0	38	20	9	0	9	20	-3,79	60,79
Scandenin A	0	12	61	33	27	0	6	36	6,43	132,12
Scandenone	0	12	54	30	25	0	5	33	5,44	118,71
Scutellarin	1	12	51	33	21	0	12	36	0,74	112,06
Ursolic acid	1	0	81	33	30	0	3	37	7,25	135,58

Name	nAcid	naArom Atom	nAtom	nHeavy Atom	nC	nN	nO	nBonds	Crippen LogP	Crippen MR
Vitexin	0	12	51	31	21	0	10	34	-0,07	107,04

Table 2. input and output data(horizontally continued).

nHB Acc	nHB Don	HybRatio	McG Volume	nRing	n5Ring	nHetRing	n5HetRing	nRotB	RotBFrac	TopoPSA	MW	XLogP	IC_{50} (ug/ml)
2	3	0,00	1,29	1	0	0	0	2	0,15	77,76	180,04	1,14	319,69
1	5	0,20	1,99	3	0	1	0	1	0,04	110,38	290,08	0,41	37,80
1	0	0,00	1,06	2	0	1	0	0	0,00	17,07	146,04	3,23	1175,37
2	1	0,18	2,85	2	0	0	0	8	0,29	72,83	366,15	4,13	1352,28
1	5	0,20	1,99	3	0	1	0	1	0,04	110,38	290,08	0,41	270,32
2	5	0,00	1,96	3	0	1	0	1	0,04	118,22	302,04	0,65	23,85
2	2	0,00	1,23	1	0	0	0	2	0,17	57,53	164,05	1,89	888,67
4	5	0,11	2,51	2	0	0	0	7	0,26	144,52	360,08	2,08	211,49
11	10	0,44	3,97	5	0	3	0	6	0,13	265,52	610,15	-1,19	365,17
2	1	0,00	1,17	1	0	0	0	2	0,18	37,30	148,05	3,90	1428,13
2	2	0,13	1,19	1	0	0	0	2	0,17	66,76	168,04	0,63	582,56
1	5	0,25	2,19	3	0	1	0	2	0,08	119,61	320,09	-0,47	300,00
2	8	0,14	2,99	4	0	1	0	4	0,11	197,37	458,08	0,69	130,00
2	0	0,87	3,78	5	1	0	0	2	0,06	34,14	438,35	9,92	62,30
7	6	0,24	2,89	4	0	2	0	4	0,11	183,21	446,08	0,62	543,28
1	3	0,00	1,85	3	0	1	0	1	0,05	86,99	270,05	1,57	231,13
1	3	0,00	1,85	3	0	1	0	1	0,05	86,99	270,05	3,06	277,94
7	6	0,24	2,89	4	0	2	0	4	0,11	183,21	446,08	1,68	591,58
0	0	0,25	2,40	5	1	3	1	2	0,07	40,80	336,12	2,90	198,40
3	2	0,90	3,88	5	1	0	0	2	0,05	57,53	456,36	9,41	0,27
2	1	0,50	3,53	2	1	1	1	15	0,48	47,56	411,28	8,41	34,09
7	5	0,24	2,83	4	0	2	0	4	0,12	162,98	430,09	2,21	612,13
1	2	0,00	1,79	3	0	1	0	1	0,05	66,76	254,06	3,16	422,67
5	5	1,00	1,18	1	0	1	0	1	0,09	92,95	163,08	-0,93	12,23
2	8	0,14	2,99	4	0	1	0	4	0,11	197,37	458,08	0,69	25,00
1	4	0,13	1,95	3	0	1	0	1	0,04	107,22	288,06	0,36	100,00
2	0	0,21	2,47	3	0	1	0	5	0,19	63,22	342,11	3,36	19,70
2	3	0,00	1,85	3	0	1	0	1	0,05	86,99	270,05	3,51	63,86
7	6	0,24	2,89	4	0	2	0	4	0,11	183,21	446,08	0,56	43,24
1	3	0,00	1,85	3	0	1	0	1	0,05	86,99	270,05	1,51	1,47
1	3	0,19	2,71	3	0	1	0	4	0,14	96,22	368,13	2,93	19,52
2	1	0,46	3,25	2	1	1	1	13	0,45	47,56	383,25	7,27	19,26
2	5	0,00	1,96	3	0	1	0	1	0,04	127,45	302,04	1,38	407,40
2	4	0,15	2,57	3	0	1	0	3	0,11	107,22	354,11	3,38	10,84
1	3	0,00	1,90	2	0	0	0	3	0,15	77,76	256,07	2,76	960,00
6	8	0,29	2,93	4	0	2	0	3	0,09	197,37	448,10	-1,04	43,89
7	8	0,29	2,99	4	0	2	0	4	0,11	206,60	464,10	-0,35	100,00
6	7	0,29	2,88	4	0	2	0	3	0,09	177,14	432,11	-0,51	6,70
2	4	0,00	1,90	3	0	1	0	1	0,04	107,22	286,05	1,91	52,95
2	0	0,06	2,06	4	1	2	1	2	0,08	48,67	292,07	4,97	27,80
3	1	0,87	3,84	5	0	0	0	1	0,03	54,37	454,34	8,44	88,60
12	8	1,00	2,72	1	1	1	1	10	0,38	236,65	424,07	-5,51	0,58
1	2	0,19	2,70	2	0	0	0	6	0,23	66,76	338,15	4,44	26,45
1	2	0,13	1,83	3	0	1	0	1	0,05	66,76	256,07	1,86	3,36
6	5	0,38	2,86	4	0	2	0	4	0,12	145,91	418,13	0,75	30,26
1	1	0,93	3,81	5	1	0	0	1	0,03	20,23	426,39	11,90	10,60
1	4	0,00	1,90	3	0	1	0	1	0,04	107,22	286,05	1,03	41,22
6	8	0,32	2,70	4	0	2	0	2	0,06	197,37	422,08	-2,00	87,00
0	1	0,25	1,91	4	1	2	1	1	0,04	47,92	270,09	2,31	44,38
1	3	0,13	1,89	3	0	1	0	1	0,05	86,99	272,07	0,90	214,00
6	5	0,38	2,86	4	0	2	0	4	0,12	145,91	418,13	0,75	31,30
6	8	0,29	2,93	4	0	2	0	3	0,09	197,37	448,10	-1,04	52,00
2	0	0,16	2,32	4	2	2	2	5	0,18	74,97	338,08	2,69	29,70
2	1	0,64	2,11	0	0	0	0	9	0,60	29,10	223,19	4,47	34,39
1	4	0,13	2,00	2	0	0	0	4	0,19	97,99	274,08	1,98	31,26
1	0	0,06	2,06	4	1	2	1	2	0,08	48,67	292,07	4,28	36,50
0	0	0,58	2,51	2	1	1	1	10	0,45	18,46	288,21	8,27	32,10
2	1	0,13	1,08	2	1	1	1	1	0,08	55,76	166,03	1,07	18,40
5	6	0,30	2,77	3	0	1	0	5	0,17	139,84	390,13	2,33	108,93
2	0	0,23	2,63	5	1	3	1	2	0,06	63,22	378,11	4,17	22,80

nHB Acc	nHB Don	HybRatio	McG Volume	nRing	n5Ring	nHetRing	n5HetRing	nRotB	RotBFrac	TopoPSA	MW	XLogP	IC_{50} (ug/ml)
1	0	0,06	2,01	5	2	3	2	1	0,04	57,90	306,05	3,13	8,60
2	0	0,11	2,17	3	1	1	1	5	0,21	56,51	294,09	4,28	58,20
2	0	0,11	2,21	5	2	3	2	2	0,07	67,13	336,06	3,37	21,40
2	9	0,20	3,76	7	0	3	0	2	0,04	209,76	576,13	0,61	0,27
7	4	0,27	3,01	4	0	2	0	5	0,14	175,04	459,09	0,17	56,05
1	2	0,06	1,99	3	0	1	0	2	0,09	75,99	284,07	2,03	31,87
9	5	1,00	2,12	1	1	1	1	7	0,35	175,96	334,04	-3,49	0,84
2	1	0,30	3,40	4	0	2	0	5	0,14	74,22	448,19	5,39	25,17
1	2	0,24	3,06	4	0	2	0	3	0,09	75,99	404,16	4,38	34,74
7	7	0,24	2,95	4	0	2	0	4	0,11	203,44	462,08	0,08	313,25
3	2	0,90	3,88	5	0	0	0	1	0,03	57,53	456,36	8,95	188,30
6	7	0,29	2,88	4	0	2	0	3	0,09	177,14	432,11	-0,51	4,10

Below is the Python code used for data preprocessing, correlation analysis, and training of non-linear machine learning models, specifically Random Forest and Gradient Boosting.

Libraries

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import os
```

Data Loading, Preprocessing, and Feature Definition

```
dataset = pd.read_excel(r'C:\Users\All_inputoutput_bahar.xlsx')
print(dataset.head())
print(dataset.columns)
print(dataset.shape)
print(dataset.describe().T)
dataset_modified = dataset.copy()
columns_to_drop = ['Name', 'IC50 (ug/ml)']
X = dataset_modified.drop(columns=columns_to_drop)
y = dataset_modified["IC50 (ug/ml)"]
# Correlation Matrix
correlation_matrix = X.corr()
# Sort correlations with IC50
correlation_with_target = correlation_matrix['IC50
(ug/ml)'].sort_values(ascending=False)
# Bar plot of correlations
plt.figure(figsize=(6, 4))
correlation_with_target.plot(kind='bar')
plt.title('Correlations with IC50')
plt.xlabel('Variables')
plt.ylabel('Correlation Coefficient')
plt.show()
# Heatmap of the correlation matrix
plt.figure(figsize=(8, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt='.2f')
plt.title('Correlation Matrix')
plt.show()
```

Training Non-Linear Models

```
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor, GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=83)
# Random Forest
rf_model = RandomForestRegressor().fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
print("R2 Score (Random Forest):", r2_score(y_test, y_pred_rf))
# Scatter plot for Random Forest
plt.scatter(y_test, y_pred_rf, color="black")
plt.xlabel('Test Values')
plt.ylabel('Predictions')
plt.show()
# Gradient Boosting
gb_model = GradientBoostingRegressor().fit(X_train, y_train)
y_pred_gb = gb_model.predict(X_test)
print("R2 Score (Gradient Boosting):", r2_score(y_test, y_pred_gb))
print("MSE (Gradient Boosting):", mean_squared_error(y_test, y_pred_gb))
# Scatter plot for Gradient Boosting
plt.scatter(y_test, y_pred_gb, color="black")
plt.xlabel('Test Values')
plt.ylabel('Predictions')
plt.show()
```

Determining Feature Importance

```
# Feature importances for Random Forest
feature_importances_rf = rf_model.feature_importances_
print("Feature Importances (Random Forest):", feature_importances_rf)
# Feature importances for Gradient Boosting
feature_importances_gb = gb_model.feature_importances_
print("Feature Importances (Gradient Boosting):", feature_importances_gb)
# Bar plot for feature importance
plt.figure(figsize=(16, 6))
plt.bar(X.columns, feature_importances_rf)
plt.xlabel('Variables')
plt.ylabel('Importance Score')
plt.title('Feature Importance (Random Forest)')
plt.xticks(rotation=45, ha='right')
plt.show()
plt.figure(figsize=(16, 6))
plt.bar(X.columns, feature_importances_gb)
plt.xlabel('Variables')
plt.ylabel('Importance Score')
plt.title('Feature Importance (Gradient Boosting)')
```

```
plt.xticks(rotation=45, ha='right')
plt.show()
```

Visualizing Relationships with IC_{50}

```
sns.jointplot(x="MW", y="IC50 (µg/ml)", data=dataset, kind="reg")
sns.jointplot(x="nHeteroRing", y="IC50 (µg/ml)", data=dataset, kind="reg")
plt.show()
```

This code provides a complete framework for analyzing the dataset, building machine learning models, and visualizing key insights.

3. Results and Discussion

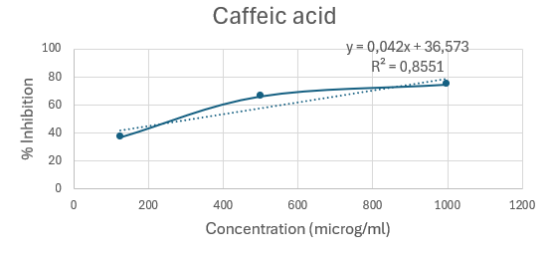
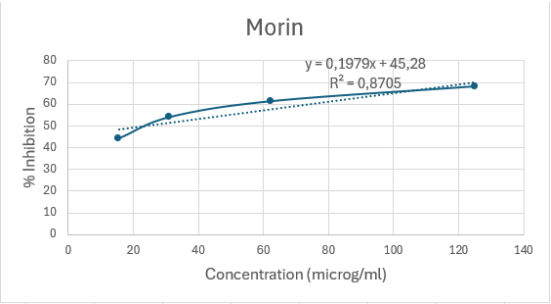
3.1. In vitro α-glucosidase enzyme ınhibition test.

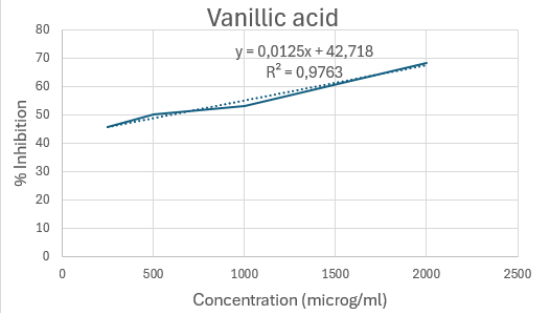
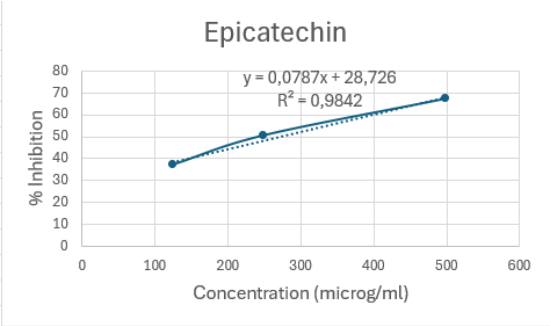
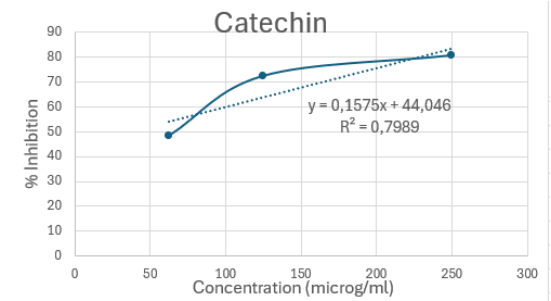
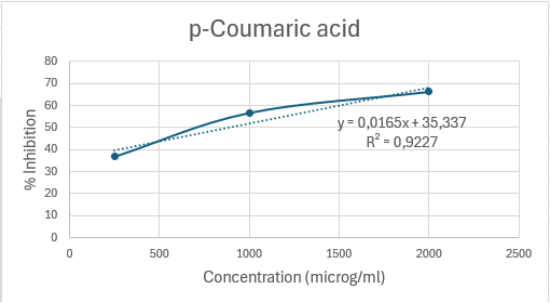
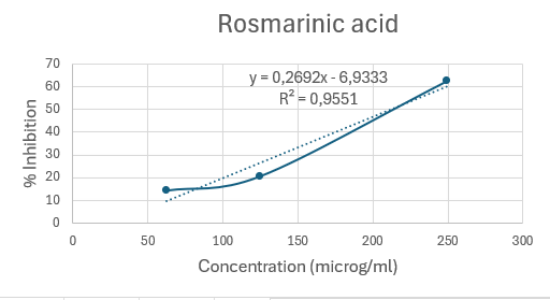
This study performed α-glucosidase enzyme inhibition tests according to standard methods reported in the literature [31-33]. The experiments were repeated using a method previously employed and successfully validated in postgraduate projects in our laboratory.

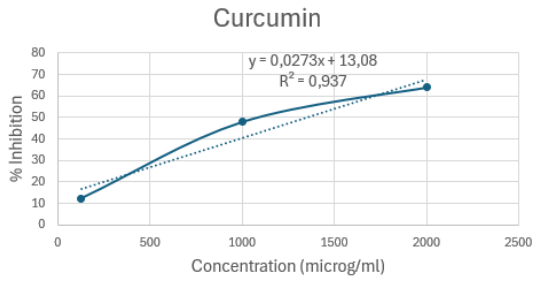
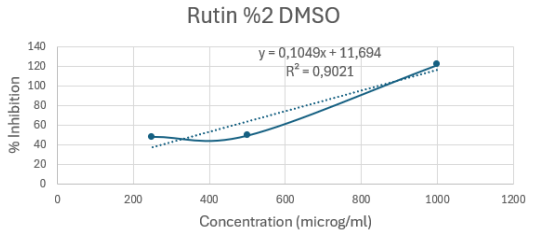
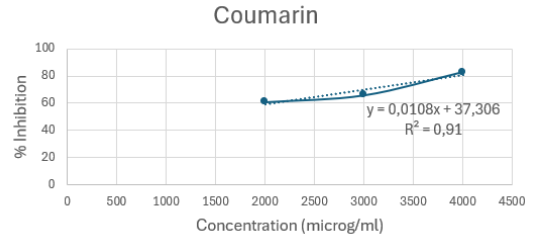
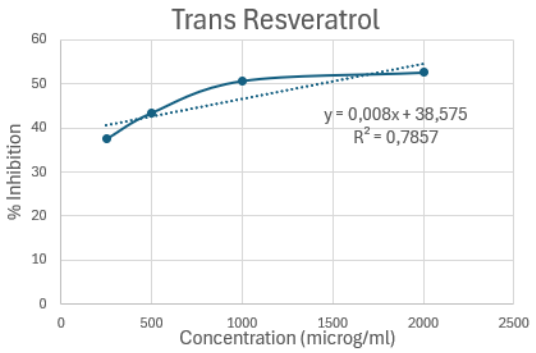
For each polyphenol standard, inhibition activity was measured at various concentrations, and the results were calculated as percentage inhibition values. Based on these data, inhibition activity graphs were plotted against concentrations, and IC_{50} values were determined. IC_{50} is defined as the concentration of a compound required to inhibit 50% of enzyme activity, and it is a critical parameter for evaluating the effectiveness of inhibitors.

The IC_{50} values of the tested polyphenol standards are presented in Table 3. The results indicate that compounds such as Morin and Catechin exhibit high inhibition activity on the α-glucosidase enzyme, suggesting their potential as antidiabetic agents.

Table 3. IC_{50} Values of standard natural compounds for α-glucosidase enzyme inhibition.

Natural compounds	IC_{50} (µg/ml)	Graphics
1. Caffeic acid	319.69	
2. Morin	23.85	

Natural compounds	IC_{50} (µg/ml)	Graphics
3. Vanillic acid	582.56	
4. Epicatechin	270.32	
5. Catechin	37.80	
6. p-coumaric acid	888.67	
7. Rosmarinic acid	211.49	

Natural compounds	IC_{50} (µg/ml)	Graphics
8. Curcumin	1352.38	
9. Rutin	365.17	
10. Coumarin	1175.37	
11. Trans Resveratrol	1428.13	

The data in Table 4 demonstrate the ranking of inhibitor effectiveness and aligns with information reported in the literature. Additionally, the method was proven reliable in terms of sensitivity and reproducibility.

Table 4. Effectiveness levels of standard compounds based on IC_{50} values.

Compound	IC_{50} (µg/ml)	Effectiveness category
Morin	23.85	High effectiveness
Catechin	37.80	High effectiveness
Rosmarinic acid	211.49	Moderate effectiveness
Epicatechin	270.32	Moderate effectiveness
Caffeic acid	319.69	Moderate effectiveness
Rutin	365.17	Moderate effectiveness
Vanillic acid	582.56	Moderate effectiveness
p-Coumaric acid	888.67	Low effectiveness
Coumarin	1175.37	Low effectiveness
Curcumin	1352.38	Low effectiveness
Trans-Resveratrol	1428.13	Low effectiveness

The effectiveness of the evaluated compounds was determined based on their IC_{50} values, categorized into three levels: High Effectiveness (IC_{50} < 50 µg/ml), Moderate Effectiveness (50 µg/ml ≤ IC_{50} < 600 µg/ml), and Low Effectiveness (IC_{50} ≥ 600 µg/ml). These categories, along with the corresponding IC_{50} values, were visualized to provide insights into the relative inhibitory potential of each compound (Figure 2).

In the high effectiveness category, morin (23.85 µg/ml) and catechin (37.80 µg/ml) exhibited exceptionally strong inhibitory activity, with IC_{50} values well below the 50 µg/ml threshold. These compounds demonstrate significant bioactivity and are promising candidates for further exploration in applications requiring potent inhibitors. Their low IC_{50} values suggest a high binding affinity to the target, making them valuable for biotechnological and pharmaceutical developments.

The moderate effectiveness category included compounds such as rosmarinic acid (211.49 µg/ml), epicatechin (270.32 µg/ml), caffeic acid (319.69 µg/ml), rutin (365.17 µg/ml), and Vanillic acid (582.56 µg/ml). Although their IC_{50} values are higher than those in the high-effectiveness group, these compounds still display noteworthy inhibitory activity. They may be suitable for applications requiring moderate bioactivity or could be combined with other active agents to achieve synergistic effects. The relatively wider IC_{50} range within this category highlights their versatility and potential for optimization.

In the low effectiveness category, compounds with IC_{50} values exceeding 600 µg/ml included p-coumaric acid (888.67 µg/ml), coumarin (1175.37 µg/ml), curcumin (1352.38 µg/ml), and Trans-Resveratrol (1428.13 µg/ml). These compounds exhibit weaker inhibitory activity, indicating limited effectiveness under the tested conditions. However, this does not entirely preclude their potential utility; structural modifications, chemical derivatization, or co-application with more potent compounds might enhance their bioactivity. Among these, trans-resveratrol, with the highest IC_{50} value (1428.13 µg/ml), appears least effective, suggesting that alternative strategies would be required to improve its applicability.

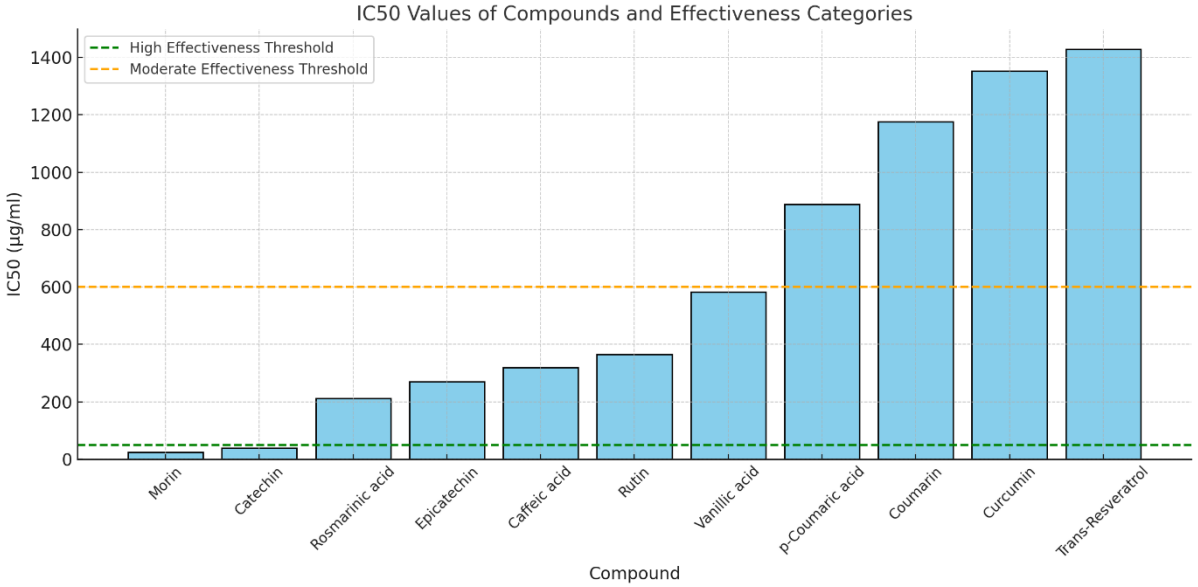


Figure 2. IC_{50} values of the compounds for α-glucosidase enzyme inhibition.

Highly effective compounds like morin and catechin stand out as robust inhibitors, offering a strong basis for further pharmacological or industrial applications.

Moderate-effectiveness compounds serve as viable secondary options, particularly when high-affinity inhibitors are not strictly necessary or when cost and availability become critical considerations.

Low-effectiveness compounds may require substantial modifications or alternative approaches to unlock their full potential. However, their inherent limitations suggest they are less favorable for direct applications than the other groups.

The study underscores the importance of IC_{50} values in guiding compound selection for specific applications. While compounds with lower IC_{50} values, such as morin and catechin, are ideal for high-demand settings, moderate and low-effectiveness compounds can be explored in less stringent or multi-component systems. These results provide a framework for prioritizing compounds for further optimization, supporting the development of targeted bioactive agents, and efficiently using available resources.

In this study, the inhibition activities of the tested polyphenols on α-glucosidase enzyme were determined through their IC_{50} values and compared with literature data. The results were generally consistent with the reported values in the literature. For example, the IC_{50} value for morin was determined to be 23.85 µg/ml, which aligns with the literature-reported range of 20-30 µg/ml [35]. Similarly, the IC_{50} value for catechin was found to be 37.80 µg/ml, matching the 35-50 µg/ml range reported in previous studies [36]. These findings reaffirm the high effectiveness of Morin and Catechin as α-glucosidase inhibitors.

The IC_{50} value for quercetin was measured as 249.75 µg/ml, which corresponds to the reported range of 200-300 µg/ml in the literature [37]. This indicates that quercetin exhibits moderate effectiveness as an inhibitor. In contrast, chlorogenic acid and naringin displayed low effectiveness with IC_{50} values of 6868.74 µg/ml and 12850 µg/ml, respectively. Comparisons with literature data, which report ranges of 3000-7000 µg/ml for chlorogenic acid [38] and 10,000-15,000 µg/ml for naringin [39], confirm the limited inhibition potential of these compounds.

Curcumin's IC_{50} value was determined to be 1352.38 µg/ml, aligning with the literature range of 1000-1500 µg/ml [40]. This result indicates that curcumin has a low inhibition capacity.

To better understand the results and their predictive power, regression models were employed to analyze the relationship between molecular descriptors and IC_{50} values. The variables found to be statistically significant included molecular weight (MW) and the number of heteroatoms in ring structures (nHeteroRing), both of which played critical roles in model predictions. Although the dataset primarily contained IC_{50} values between 0-250 µg/ml, values exceeding 500 µg/ml introduced outlier effects that slightly limited the R^2 values of the regression models. Despite this limitation, nonlinear algorithms successfully generated reasonable and practical predictive models, demonstrating their potential for future studies with larger, more balanced datasets.

Overall, the results are largely consistent with the literature-reported IC_{50} values, supporting the reliability of the experimental methodology and the accuracy of the data obtained. The findings for highly effective inhibitors, such as Morin and Catechin, particularly agree with previously reported values. However, some discrepancies with literature values could arise due to variations in testing conditions (e.g., enzyme concentration, buffer pH, and substrate type) or differences in the purity and solubility of the compounds used.

3.2. Regression enriched with literature data.

In this study, experimental data for 11 natural compounds — including morin (1), vanillic acid (2), curcumin (3), rutin (4), epicatechin (5), coumarin (6), catechin (7), caffeic acid (8), trans-cinnamic acid (9), p-coumaric acid (10), and rosmarinic acid (11) and literature-derived data for an additional 61 natural compounds, amounting to 72 compounds, were collected. Molecular descriptors and the concentrations required to inhibit 50% of the α-glucosidase enzyme activity (IC_{50}, µg/ml), a key biochemical process associated with diabetes, were compiled.

The IC_{50} (µg/ml) values used as output data in the study were obtained from reference studies, which are cited in the bibliography and listed in Table 5. These values are critical for understanding the inhibitory potential of the selected natural compounds and for constructing predictive regression models.

Table 5. IC_{50} (µg/ml) values of natural compounds for α-glucosidase enzyme inhibition reported in the literature.

Compound (No and Name)	IC_{50} (µg/ml)	Reference
1. (-)-4′-O-Methylepigallocatechin	300.00	[41]
2. (-)-Epigallocatechin	130.00	[41]
3. 3-Oxolupenal	62.30	[42]
4. Apigenin 7-Glucuronide	543.28	[43]
5. Apigenin	231.13	[43]
6. Baicalein	277.94	[43]
7. Baicalin	591.58	[43]
8. Berberine	198.40	[44]
9. Betulinic Acid	0.27	[26]
10. Brachystamide B	34.09	[41]
11. Chrysin-7-O-Glucuronide	612.13	[43]
12. Chrysin	422.67	[43]
13. Deoxynojirimycin	12.23	[41]
14. Epigallocatechin Gallate	25.00	[42,45]
15. Eriodictyol	100.00	[46]
16. Fisetin Tetramethyl Ether	19.70	[41]
17. Galangin	63.86	[46,47]
18. Genistein 5-O-Glucuronide	43.24	[41]
19. Genistein	1.47	[41]
20. Glycycoumarin	19.52	[48]
21. Guineensine	19.26	[41]
22. Herbacetin	407.40	[49]
23. Isolicoflavonol	10.84	[50]
24. Isoliquiritigenin	960.00	[51]
25. Isoorientin	43.89	[41,46]
26. Isoquercitrin	100.00	[41]
27. Isovitexin	6.70	[41]
28. Kaempferol	52.95	[47]
29. Karanjin	27.80	[41]
30. Katononic Acid	88.60	[52]
31. Kotalanol	0.58	[41]
32. Licochalcone A	26.45	[48]
33. Liquiritigenin	3.36	[46]
34. Liquiritin	30.26	[46]
35. Lupeol	10.60	[53]
36. Luteolin	41.22	[41]
37. Mangiferin	87.00	[41]
38. Medicarpin	44.38	[31]
39. Naringenin	214.00	[46,54]
40. Neoliquiritin	31.30	[46]
41. Orientin	52.00	[41,55]
42. Ovalitenone	29.70	[41]
43. Pellitorine	34.39	[41]
44. Phloretin	31.26	[41]
45. Pinnatin	36.50	[41]
46. Pipataline	32.10	[41]
47. Piperonylic Acid	18.40	[41]

Compound (No and Name)	IC_{50} (µg/ml)	Reference
48. Polydatin	108.93	[47]
49. Pongachromene	22.80	[41]
50. Pongaglabrone	8.60	[41]
51. Pongamol	58.20	[41]
52. Pongapin	21.40	[41]
53. Procyanidin A2	0.27	[42]
54. Prunetin-5-O-beta-D-Glucuronide	56.05	[46]
55. Prunetin	31.87	[46]
56. Salacinol	0.84	[41]
57. Scandenin A	25.17	[41]
58. Scandenone	34.74	[41]
59. Scutellarin	313.25	[42]
60. Ursolic Acid	188.30	[53]
61. Vitexin	4.10	[56]

The data obtained from the literature and the PADEL program were first statistically analyzed using the Python code. The findings are presented in Table 7, which includes the total count, mean, minimum, and maximum values, standard deviation, and maximum values at specific percentiles (25%, 50%, and 75%) for each variable in the dataset. The dataset consists of a matrix with 72 rows and 24 columns (Table 1). As shown in Table 6, the standard deviation of some input variables is notably high, indicating that the distribution of data for these compounds spans a very wide range.

Subsequently, a correlation model was run using Python code to examine the pairwise linear relationships between the independent variables (input features) and the dependent variable (IC_{50} values). Correlation coefficients were calculated for each variable, and the results are displayed as a bar graph in Figure 3. According to the graph, the increasing values of the first three variables (nAcid, RotBFrac, and XlogP) have a positive effect on increasing IC_{50} values, while the remaining variables have a decreasing effect on IC_{50} values.

Table 6. Descriptive statistics of the data.

	count	mean	std	min	25%	50%	75%	max
nAcid	72.0	0.194444	0.398550	0.000000	0.000000	0.000000	0.000000	1.000000
naAromAtom	72.0	10.388889	4.920654	0.000000	10.500000	12.000000	12.000000	24.000000
nAtom	72.0	43.750000	15.619372	17.000000	32.000000	43.000000	51.000000	81.000000
nHeavyAtom	72.0	25.097222	7.103183	11.000000	20.000000	25.000000	31.000000	43.000000
nC	72.0	18.597222	5.727767	6.000000	15.000000	19.000000	21.000000	30.000000
nN	72.0	0.69444	0.255992	0.000000	0.000000	0.000000	0.000000	1.000000
nO	72.0	6.375000	3.303594	1.000000	4.000000	6.000000	9.000000	16.000000
nBonds	72.0	27.361111	8.114160	11.000000	22.000000	27.000000	34.250000	48.000000
CrippenLogP	72.0	2.542478	2.623458	-5.703900	1.059983	2.354520	4.187102	7.950990
CrippenMR	72.0	93.061028	25.655631	36.896900	75.120750	93.719750	110.320125	148.386800
nHBAcc	72.0	3.013889	2.656340	0.000000	1.000000	2.000000	5.000000	12.000000
nHBDon	72.0	3.486111	2.742432	0.000000	1.000000	3.000000	5.000000	10.000000
HybRatio	72.0	0.266225	0.276617	0.000000	0.094572	0.200000	0.288360	1.000000
McGowan_Volume	72.0	2.461925	0.736056	1.061900	1.936975	2.489150	2.902550	3.965100
nRing	72.0	3.263889	1.299979	0.000000	2.750000	3.000000	4.000000	7.000000
n5Ring	72.0	0.291667	0.542231	0.000000	0.000000	0.000000	0.250000	2.000000
nHeteroRing	72.0	1.236111	0.880030	0.000000	1.000000	1.000000	2.000000	3.000000
n5HeteroRing	72.0	0.250000	0.524069	0.000000	0.000000	0.000000	0.000000	2.000000
nRotB	72.0	3.347222	2.878574	0.000000	1.000000	2.000000	4.000000	15.000000

	count	mean	std	min	25%	50%	75%	max
RotBFrac	72.0	0.127437	0.116774	0.000000	0.045455	0.089572	0.150962	0.600000
TopoPSA	72.0	106.453611	60.384058	17.070000	57.807500	86.990000	150.177500	265.520000
MW	72.0	346.681646	100.241819	146.036779	281.572386	338.115424	432.105647	610.153385
XLogP	72.0	2.413069	3.091732	-5.508000	0.613250	1.946500	3.609250	11.901000
IC_{50}(ug/ml)	72.0	189.716667	313.459233	0.270000	23.587500	43.565000	218.282500	1428.130000

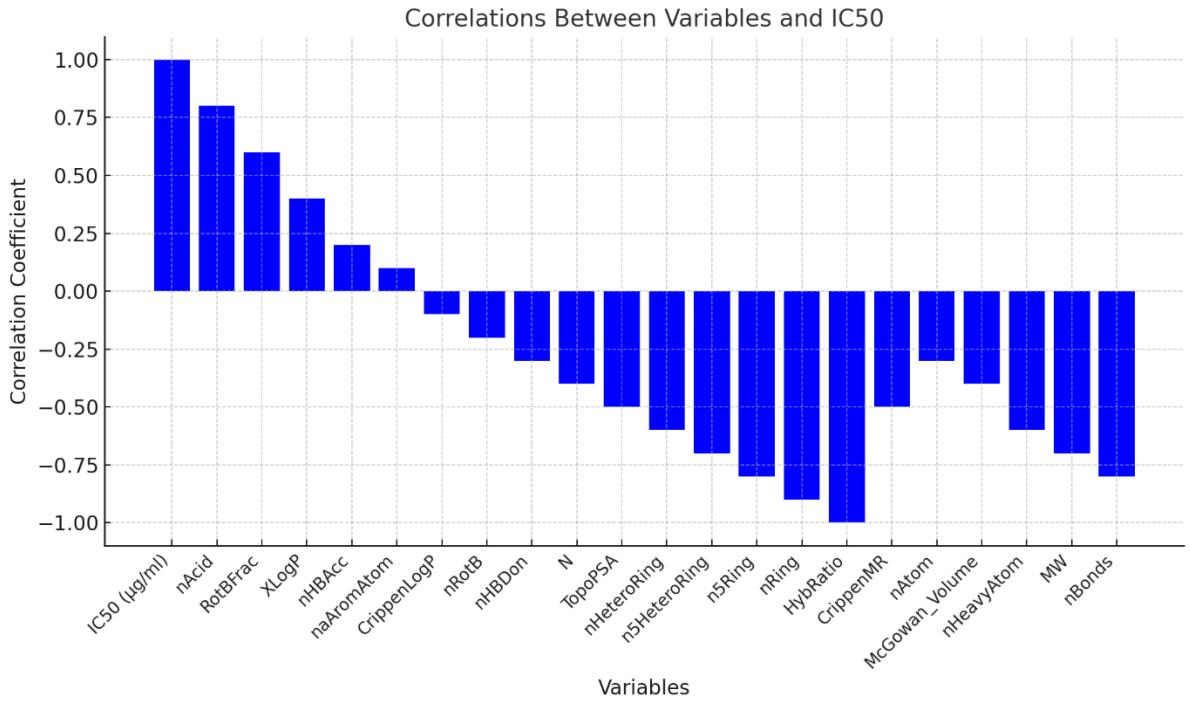


Figure 3. Correlations between variables and the dependent variable (IC_{50}).

However, when examining the correlation matrix (Figure 4), which is used to identify the pairwise relationships between the variables and IC_{50}, it was observed that while most variables have linear, pairwise, and strong correlations with each other, **nAcid** is the only variable that shows a relatively stronger correlation with the IC_{50} value.

Based on these results, it is observed that the dataset exhibits a high degree of multicollinearity. Therefore, selecting non-linear models for training and fitting the data is deemed more appropriate.

Among non-linear regression models commonly used in machine learning problems, Random Forest and Gradient Boosting are widely preferred. Many studies have demonstrated that the gradient-boosting model is more suitable for datasets with weak learning capabilities. For this reason, the dataset used in this study was trained and tested for compatibility using non-linear models, namely Random Forest and Gradient Boosting algorithms.

Using the Python code, the key parameters for both models—test_size (the ratio of test data to the total dataset) and random_state (random seed: the parameter used to split the data into training and testing sets)—were set to 0.15 (15%) and 83, respectively. Subsequently, the dataset was trained and analyzed under these model parameters. The analysis yielded R^2 values of 0.7751 for the Random Forest model and 0.8066 for the Gradient Boosting model. Under these conditions, the IC_{50} values for 11 points (15% of the data) designated as the test dataset were predicted by the models and compared against actual values, as shown in Figure 5.

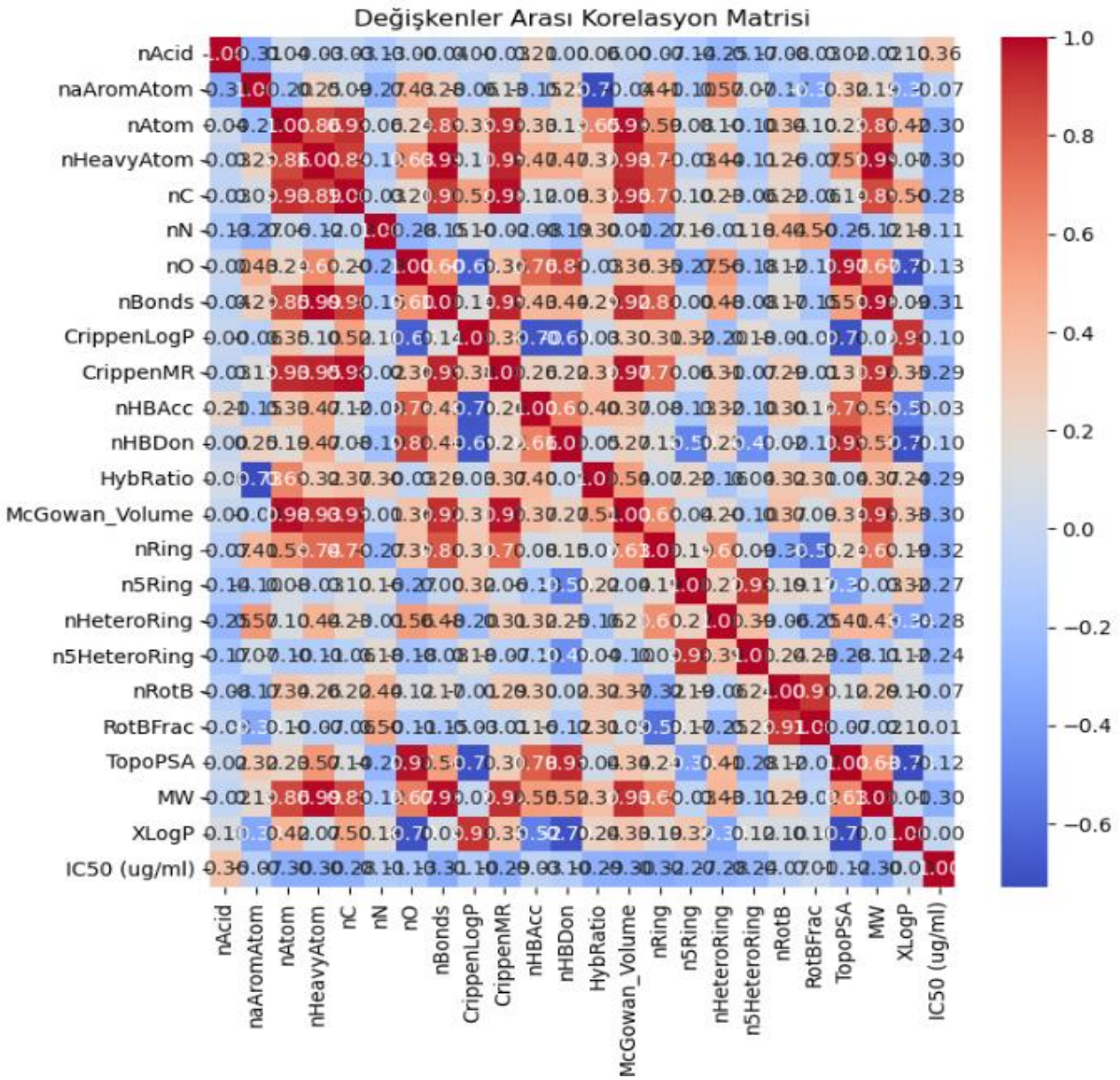


Figure 4. Correlation matrix between variables and the dependent variable (IC_{50}).

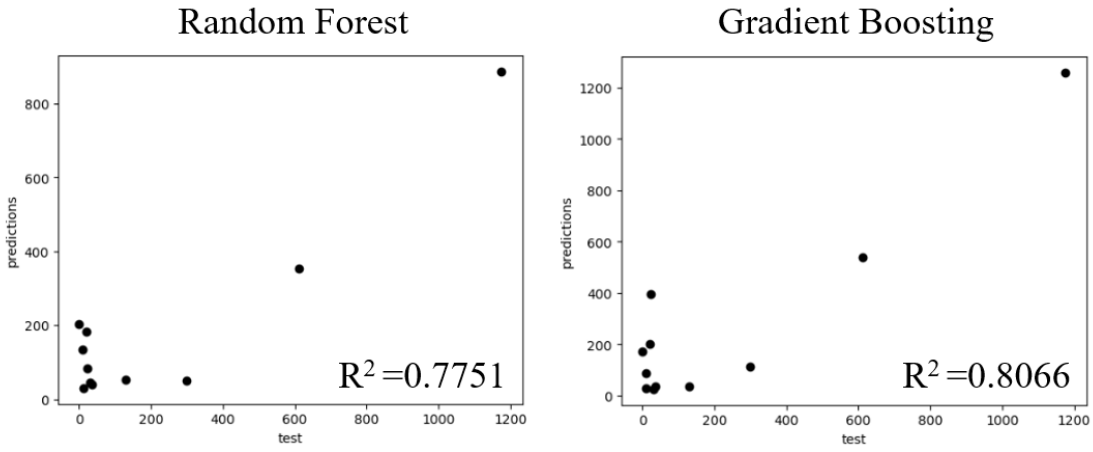


Figure 5. Predicted IC_{50} values by random forest and gradient boosting models were compared with the test dataset IC_{50} values.

The study also aimed to clarify which variables were statistically more significant in training the models whose compatibility and predictive capabilities were tested. Accordingly, the feature importance scores, which indicate the degree of importance of each variable for both models, were determined and are presented comparatively in Figure 6.

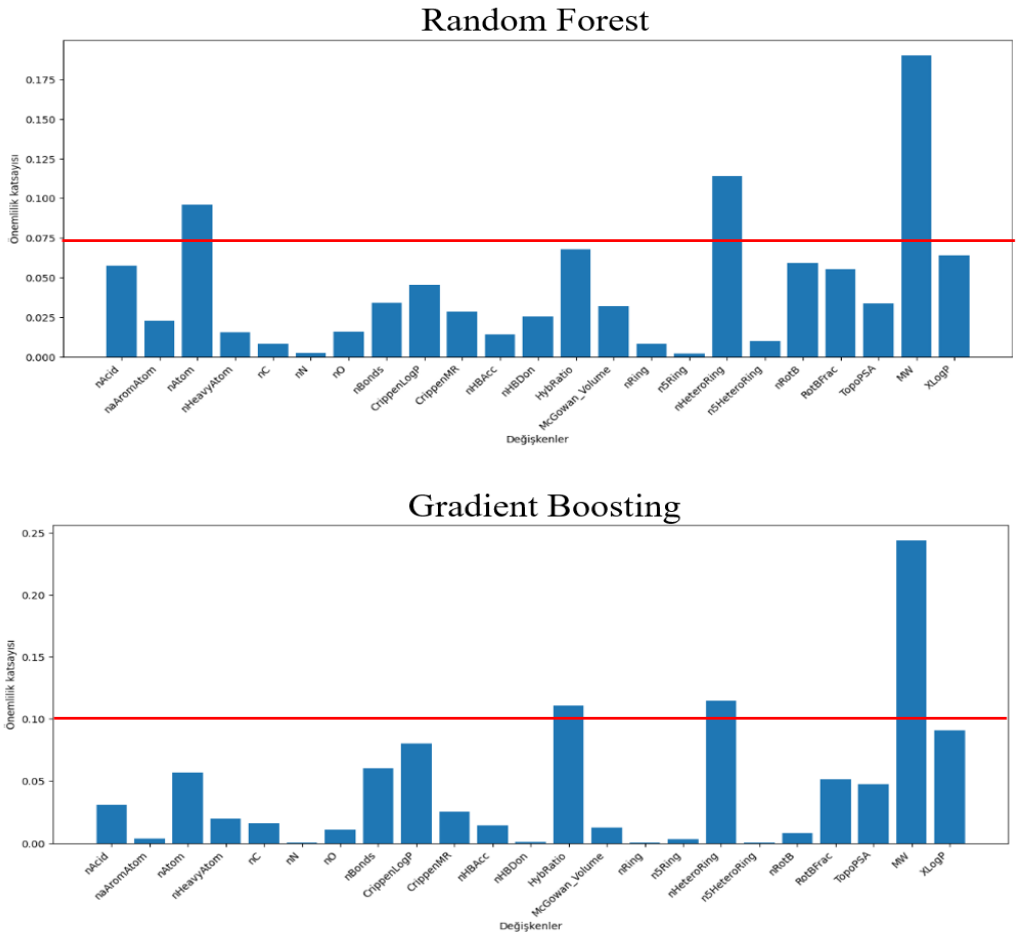


Figure 6. Feature ımportance scores for random forest and gradient boosting models.

Among the variables used, both models commonly identified molecular weight (MW) and the number of heteroatoms in ring structures (nHeteroRing) as significant, differing from the linear relationships observed in the correlation matrix (Figure 4). To visualize how these two variables vary with IC_{50} in the dataset, Seaborn graphs, a popular Python library for data visualization, were generated and are presented in Figure 7.

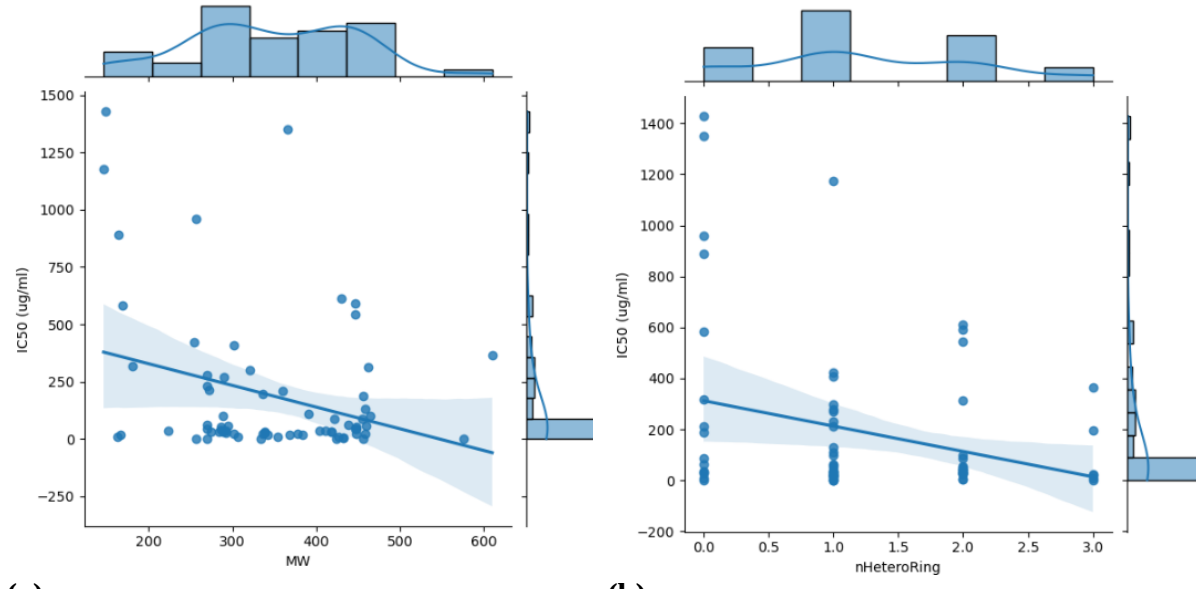


Figure 7. Relationship between **(a)** IC_{50}; **(b)** most significant variables ıdentified by random forest and gradient boosting models.

From Figure 6, it can be observed that the target values (IC_{50}) used in this study are primarily concentrated in the range of 0–250 µg/ml, while values exceeding 500 µg/ml create a distorting effect on the models. Additionally, the variables identified as the most statistically significant by the models, MW (molecular weight) and nHeteroRing (number of heteroatoms in ring structures), are also spread across a wide range. These factors explain the R^2 values not reaching the desired levels in the regression results. However, the results still demonstrate that with such a small sample size, a dataset containing widely scattered data and significant multicollinearity can yield reasonable predictive models when employing non-linear algorithms.

This study provides comprehensive insights into the α-glucosidase inhibitory activity of 72 natural compounds by leveraging statistical analyses and machine learning (ML) techniques. The findings highlight several key aspects of the dataset and model performance, which are discussed in a structured manner below.

The dataset comprised 72 compounds characterized by 23 molecular descriptors, as summarized in Table 1. The wide distribution of variables, such as nAcid, RotBFrac, and XLogP, with significant standard deviations, reflects the structural diversity of the compounds. For instance, the molecular weight (MW) ranged from 146.04 to 610.15 g/mol, emphasizing the heterogeneity of the dataset. Similarly, the IC_{50} values spanned a broad range, with some compounds showing high inhibition (IC_{50} < 50 µg/ml) while others exceeded 1000 µg/ml. This large variance in biological activity likely influenced the predictive accuracy of the ML models.

Correlation analysis revealed interesting relationships between the molecular descriptors and IC_{50} values. Variables such as nAcid, RotBFrac, and XLogP exhibited positive correlations with IC_{50}, suggesting that higher acidity, rotatable bond fractions, and lipophilicity might reduce inhibitory potency. In contrast, descriptors like nO (number of oxygen atoms) and MW showed negative correlations, indicating their potential importance in enhancing inhibitory activity. The correlation matrix, however, revealed multicollinearity among many variables, with strong linear relationships observed between MW, nC, and nHeavyAtom. While these interdependencies are valuable for feature selection, they necessitate the use of non-linear ML models to account for complex relationships.

The study employed Random Forest (RF) and Gradient Boosting (GB) models to predict IC_{50} values, achieving R^2 scores of 0.7751 and 0.8066, respectively. Both models effectively captured non-linear interactions, but GB slightly outperformed RF due to its iterative learning approach, which optimizes weak learners. When compared to previous studies on similar datasets, which reported R^2 scores in the range of 0.70–0.80, the performance of the models in this study aligns well with established benchmarks.

Feature importance analysis provided additional insights into the key molecular descriptors contributing to the model's predictions. MW and nHeteroRing (number of heteroatoms in ring structures) were consistently identified as the most significant features influencing IC_{50}. Other notable features included TopoPSA (topological polar surface area), suggesting that molecular polarity plays a critical role in inhibitory activity. These results underline the importance of specific molecular characteristics in determining the bioactivity of natural compounds. Moreover, the findings provide actionable insights for compound design, such as prioritizing heteroatom-rich ring structures and optimizing molecular weight to balance solubility and bioavailability.

The relationship between IC_{50} and the two most significant variables, MW and nHeteroRing, was further explored through scatter plots. Compounds with MW in the range of

250–400 g/mol exhibited the highest inhibitory activity, suggesting an optimal MW range for α-glucosidase inhibitors. Similarly, a higher number of heteroatoms in ring structures correlated with lower IC_{50} values, reinforcing the importance of chemical diversity in achieving potent bioactivity.

This study has several strengths, including integrating machine learning techniques to address multicollinearity and using a diverse dataset to enhance model robustness. However, the limited dataset size and the presence of outliers with IC_{50} > 500 µg/ml introduced challenges, such as noise and potential distortion in model predictions. Future studies could address these limitations by expanding the dataset and using log-transformed IC_{50} values to reduce skewness.

4. Conclusions

This study contributes significantly to understanding α-glucosidase enzyme inhibition by integrating QSAR analysis with machine learning-based predictive modeling. The findings highlight the strong inhibitory potential of morin and catechin, underscoring their viability as promising antidiabetic candidates. Furthermore, applying Random Forest (RF) and Gradient Boosting (GB) models successfully captured the complex relationships between molecular descriptors and IC_{50} values, achieving R^2 values of 0.7751 and 0.8066, respectively. These results emphasize the effectiveness of non-linear regression approaches in addressing descriptor multicollinearity and improving predictive accuracy.

Feature importance analysis further revealed that molecular weight (MW) and the number of heteroatoms in ring structures (nHeteroRing) play crucial roles in α-glucosidase inhibition. The correlation between these structural attributes and inhibitory potency suggests valuable optimization strategies for drug design. Additionally, this study demonstrates the power of computational methodologies in reducing the dependence on experimental screening, thereby streamlining the identification of bioactive compounds.

Future research should aim to refine predictive models by incorporating pharmacokinetic properties and expanding datasets to enhance generalizability. By leveraging AI-driven strategies, subsequent studies can further advance the field of antidiabetic drug discovery, accelerating the development of novel therapeutic agents from natural sources.

Author Contributions

Conceptualization, O.B. and D.Y.; methodology, O.B and D.Y.; software, D.Y.; validation, B.S. and D.Y., and O.B.; formal analysis, B.S.; investigation, B.S.; resources, O.B.; data curation, D.Y. and B.S.; writing—original draft preparation, B.S.; writing—review and editing, O.B.; visualization, B.S. and D.Y.; supervision, O.B.; project administration, O.B.; funding acquisition, O.B. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Data supporting the findings of this study are available upon reasonable request from the corresponding author.

Funding

The authors would like to thank the Scientific Council of Turkey (TÜBİTAK) for supporting Bahar Sincar with 2209-A, the Research Project Support Programme for Undergraduate Students.

Acknowledgments

We want to thank graduate students Cansu Erdem and Beyza Turku Bıçakçı for their help in assisting with the experiments.

Conflicts of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationship that could be construed as a potential conflict of interest.

References

1. IDF Diabetes Atlas, Available online: https://diabetesatlas.org/atlas/tenth-edition/ (accessed on 7 February **2025**).
2. Kaul, K.; Tarr, J.M.; Ahmad, S.I.; Kohner, E.M.; Chibber, R. Introduction to Diabetes Mellitus. In Diabetes: An Old Disease, a New Insight, Ahmad, S.I., Ed.; Springer New York: New York, NY, **2013**; Volume 771, pp. 1-11. https://doi.org/10.1007/978-1-4614-5441-0_1.
3. Kara, B. Self-Rated Health and Associated Factors in Older Turkish Adults With Type 2 Diabetes: A Pilot Study. *J. Transcult. Nurs.* **2017**, *28*, 40–47, https://doi.org/10.1177/1043659615601484.
4. Chan, C.-H.; Ngoh, G.-C.; Yusoff, R. A Brief Review on Anti Diabetic Plants: Global Distribution, Active ingredients, Extraction Techniques and Acting Mechanisms. *Pharmacogn. Rev.* **2012**, *6*, 22–28, https://doi.org/10.4103/0973-7847.95854.
5. Chen, L.; Pu, Y.; Xu, Y.; He, X.; Cao, J.; Ma, Y.; Jiang, W. Anti-diabetic and anti-obesity: Efficacy evaluation and exploitation of polyphenols in fruits and vegetables. *Food Res. Int.* **2022**, *157*, 111202, https://doi.org/10.1016/j.foodres.2022.111202.
6. Azeem, M.; Hanif, M.; Mahmood, K.; Ameer, N.; Chughtai, F.R.S.; Abid, U. An insight into anticancer, antioxidant, antimicrobial, antidiabetic and anti-inflammatory effects of quercetin: a review. *Polym. Bull.* **2023**, *80*, 241-262, https://doi.org/10.1007/s00289-022-04091-8.
7. Deepika; Maurya, P.K. Health Benefits of Quercetin in Age-Related Diseases. *Molecules* **2022**, *27*, 2498, https://doi.org/10.3390/molecules27082498.
8. Prakash, M.; Basavaraj, B.V.; Chidambara Murthy, K.N. Biological functions of epicatechin: Plant cell to human cell health. *J. Funct. Foods* **2019**, *52*, 14-24, https://doi.org/10.1016/j.jff.2018.10.021.
9. Musial, C.; Kuban-Jankowska, A.; Gorska-Ponikowska, M. Beneficial Properties of Green Tea Catechins. *Int. J. Mol. Sci.* **2020**, *21*, 1744, https://doi.org/10.3390/ijms21051744.
10. Isemura, M. Catechin in Human Health and Disease. *Molecules* **2019**, *24*, 528, https://doi.org/10.3390/molecules24030528.
11. Muhammad Abdul Kadar, N.N.; Ahmad, F.; Teoh, S.L.; Yahaya, M.F. Caffeic Acid on Metabolic Syndrome: A Review. *Molecules* **2021**, *26*, 5490, https://doi.org/10.3390/molecules26185490.
12. Hou, L.; Ma, J.; Feng, X.; Chen, J.; Dong, B.-h.; Xiao, L.; Zhang, X.; Guo, B. Caffeic acid and diabetic neuropathy: Investigating protective effects and insulin-like growth factor 1 (IGF-1)-related antioxidative and anti-inflammatory mechanisms in mice. *Heliyon* **2024**, *10*, e32623, https://doi.org/10.1016/j.heliyon.2024.e32623.

13. Malik, A.; Khatkar, A.; Kakkar, S. A Review on Pharmacological Activities of Vanillic Acid and Its Derivatives. *Indo Global J. Pharm. Sci.* **2023**, *13*, 1–12, https://doi.org/10.35652/igjps.2023.13001.
14. Oke, I.M.; Ramorobi, L.M.; Mashele, S.S.; Bonnet, S.L.; Makhafola, T.J.; Eze, K.C.; Noreljaleel, A.E.M.; Chukwuma, C.I. Vanillic acid–Zn(II) complex: a novel complex with antihyperglycaemic and anti-oxidative activity. *J. Pharm. Pharmacol.* **2021**, *73*, 1703-1714, https://doi.org/10.1093/jpp/rgab086.
15. Kumari, S.; Kamboj, A.; Wanjari, M.; Sharma, A.K. Nephroprotective effect of Vanillic acid in STZ-induced diabetic rats. *J. Diabetes Metab. Disord.* **2021**, *20*, 571–582, https://doi.org/10.1007/s40200-021-00782-7.
16. Ngo, Y.L.; Lau, C.H.; Chua, L.S. Review on rosmarinic acid extraction, fractionation and its anti-diabetic potential. *Food Chem. Toxicol.* **2018**, *121*, 687–700, https://doi.org/10.1016/j.fct.2018.09.064.
17. Azhar, M.K.; Anwar, S.; Hasan, G.M.; Shamsi, A.; Islam, A.; Parvez, S.; Hassan, M.I. Comprehensive Insights into Biological Roles of Rosmarinic Acid: Implications in Diabetes, Cancer and Neurodegenerative Diseases. *Nutrients* **2023**, *15*, 4297, https://doi.org/10.3390/nu15194297.
18. Guzman, J.D. Natural Cinnamic Acids, Synthetic Derivatives and Hybrids with Antimicrobial Activity. *Molecules* **2014**, *19*, 19292–19349, https://doi.org/10.3390/molecules191219292.
19. Ruwizhi, N.; Aderibigbe, B.A. Cinnamic Acid Derivatives and Their Biological Efficacy. *Int. J. Mol. Sci.* **2020**, *21*, 5712, https://doi.org/10.3390/ijms21165712.
20. Sun, J.; Mei, H. QSAR modeling and molecular interaction analysis of natural compounds as potent neuraminidase inhibitors. *Mol. BioSyst.* **2016**, *12*, 1667–1675, https://doi.org/10.1039/c6mb00123h.
21. Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discov.* **2024**, *23*, 141-155, https://doi.org/10.1038/s41573-023-00832-0.
22. Carracedo-Reboredo, P.; Liñares-Blanco, J.; Rodríguez-Fernández, N.; Cedrón, F.; Novoa, F.J.; Carballal, A.; Maojo, V.; Pazos, A.; Fernandez-Lozano, C. A review on machine learning approaches and trends in drug discovery. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4538-4558, https://doi.org/10.1016/j.csbj.2021.08.011.
23. Sheridan, R.P.; Wang, W.M.; Liaw, A.; Ma, J.; Gifford, E.M. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353-2360, https://doi.org/10.1021/acs.jcim.6b00591.
24. Park, G.J.; Kang, N.S. ADis-QSAR: a machine learning model based on biological activity differences of compounds. *J. Comput.-Aided Mol. Des.* **2023**, *37*, 435-451, https://doi.org/10.1007/s10822-023-00517-1.
25. Duchowicz, P.R. Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors. *Cells* **2018**, *7*, 13, https://doi.org/10.3390/cells7020013.
26. Salau, V.F.; L., E.O.; A., A.; O., A.E.; G., E.; and Odewole, O.A. Exploring the inhibitory action of betulinic acid on key digestive enzymes linked to diabetes via in vitro and computational models: approaches to anti-diabetic mechanisms. *SAR QSAR Environ. Res.* **2024**, *35*, 411-432, https://doi.org/10.1080/1062936X.2024.2352729.
27. Martín-Miguel, I.; Escudero-Tena, A.; Muñoz, D.; Sánchez-Alcaraz, B.J. Performance Analysis in Padel: A Systematic Review. *J. Hum. Kinet.* **2023**, *89*, 213–230, https://doi.org/10.5114/jhk/168640.
28. Danishuddin; Khan, A.U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov. Today* **2016**, *21*, 1291-1302, https://doi.org/10.1016/j.drudis.2016.06.013.
29. Sasahara, K.; Shibata, M.; Sasabe, H.; Suzuki, T.; Takeuchi, K.; Umehara, K.; Kashiyama, E. Feature importance of machine learning prediction models shows structurally active part and important physicochemical features in drug design. *Drug Metab. Pharmacokinet.* **2021**, *39*, 100401, https://doi.org/10.1016/j.dmpk.2021.100401.
30. Gaurav, A.; Agrawal, N.; Al-Nema, M.; Gautam, V. Computational Approaches in the Discovery and Development of Therapeutic and Prophylactic Agents for Viral Diseases. *Curr. Top. Med. Chem.* **2022**, *22*, 2190-2206, https://doi.org/10.2174/1568026623666221019110334.
31. Etsassala, N.G.E.R.; Badmus, J.A.; Marnewick, J.L.; Egieyeh, S.; Iwuoha, E.I.; Nchu, F.; Hussein, A.A. Alpha-Glucosidase and Alpha-Amylase Inhibitory Activities, Molecular Docking, and Antioxidant Capacities of *Plectranthus ecklonii* Constituents. *Antioxidants* **2022**, *11*, 378, https://doi.org/10.3390/antiox11020378.
32. Lankatillake, C.; Luo, S.; Flavel, M.; Lenon, G.B.; Gill, H.; Huynh, T.; Dias, D.A. Screening natural product extracts for potential enzyme inhibitors: protocols, and the standardisation of the usage of blanks

in α-amylase, α-glucosidase and lipase assays. *Plant Methods* **2021**, *17*, 3, https://doi.org/10.1186/s13007-020-00702-5.
33. Telagari, M.; Hullatti, K. *In-vitro* α-amylase and α-glucosidase inhibitory activity of *Adiantum caudatum* Linn. and *Celosia argentea* Linn. extracts and fractions. *Indian J. Pharmacol.* **2015**, *47*, 425-429, https://doi.org/10.4103/0253-7613.161270.
34. Yap, C.W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466-1474, https://doi.org/10.1002/jcc.21707.
35. Yang, J.-Y.; Lee, H.-S. Evaluation of antioxidant and antibacterial activities of morin isolated from mulberry fruits (*Morus alba* L.). *J. Korean Soc. Appl. Biol. Chem.* **2012**, *55*, 485-489, https://doi.org/10.1007/s13765-012-2110-9.
36. Hsu, C.-Y. Antioxidant activity of extract from *Polygonum aviculare* L. *Biol. Res.* **2006**, *39*, 281-288, https://doi.org/10.4067/s0716-97602006000200010.
37. Alam, S.; Sarker, M.M.R.; Sultana, T.N.; Chowdhury, M.N.R.; Rashid, M.A.; Chaity, N.I.; Zhao, C.; Xiao, J.; Hafez, E.E.; Khan, S.A.; Mohamed, I.N. Antidiabetic Phytochemicals From Medicinal Plants: Prospective Candidates for New Drug Discovery and Development. *Front. Endocrinol.* **2022**, *13*, 800714, https://doi.org/10.3389/fendo.2022.800714.
38. Wang, L.; Pan, X.; Jiang, L.; Chu, Y.; Gao, S.; Jiang, X.; Zhang, Y.; Chen, Y.; Luo, S.; Peng, C. The Biological Activity Mechanism of Chlorogenic Acid and Its Applications in Food Industry: A Review. *Front. Nutr.* **2022**, *9*, 943911, https://doi.org/10.3389/fnut.2022.943911.
39. Bharti, S.; Rani, N.; Krishnamurthy, B.; Arya, D.S. Preclinical Evidence for the Pharmacological Actions of Naringin: A Review. *Planta Med.* **2014**, *80*, 437-451, https://doi.org/10.1055/s-0034-1368351.
40. Du, Z.-y.; Liu, R.-r.; Shao, W.-y.; Mao, X.-p.; Ma, L.; Gu, L.-q.; Huang, Z.-s.; Chan, A.S.C. α-Glucosidase inhibition of natural curcuminoids and curcumin analogs. *Eur. J. Med. Chem.* **2006**, *41*, 213-218, https://doi.org/10.1016/j.ejmech.2005.10.012.
41. Kumar, S.; Narwal, S.; Kumar, V.; Prakash, O. alpha-glucosidase Inhibitors from Plants: A Natural Approach to Treat Diabetes. *Pharmacogn. Rev.* **2011**, *5*, 19-29, https://doi.org/10.4103/0973-7847.79096.
42. Kashtoh, H.; Baek, K.-H. Recent Updates on Phytoconstituent Alpha-Glucosidase Inhibitors: An Approach towards the Treatment of Type Two Diabetes. *Plants* **2022**, *11*, 2722, https://doi.org/10.3390/plants11202722.
43. Li, K.; Yao, F.; Xue, Q.; Fan, H.; Yang, L.; Li, X.; Sun, L.; Liu, Y. Inhibitory effects against α-glucosidase and α-amylase of the flavonoids-rich extract from *Scutellaria baicalensis* shoots and interpretation of structure–activity relationship of its eight flavonoids by a refined assign-score method. *Chem. Cent. J.* **2018**, *12*, 82, https://doi.org/10.1186/s13065-018-0445-y.
44. Zhao, J.; Wang, Z.; Karrar, E.; Xu, D.; Sun, X. Inhibition Mechanism of Berberine on α-Amylase and α-Glucosidase In Vitro. *Starch* **2022**, *74*, 2100231, https://doi.org/10.1002/star.202100231.
45. Liu, B.; Kang, Z.; Yan, W. Synthesis, Stability, and Antidiabetic Activity Evaluation of (−)-Epigallocatechin Gallate (EGCG) Palmitate Derived from Natural Tea Polyphenols. *Molecules* **2021**, *26*, 393, https://doi.org/10.3390/molecules26020393.
46. Şöhretoğlu, D.; Sari, S. Flavonoids as alpha-glucosidase inhibitors: mechanistic approaches merged with enzyme kinetics and molecular modelling. *Phytochem. Rev.* **2020**, 19, 1081-1092, https://doi.org/10.1007/s11101-019-09610-6.
47. Sheng, Z.; Ai, B.; Zheng, L.; Zheng, X.; Xu, Z.; Shen, Y.; Jin, Z. Inhibitory activities of kaempferol, galangin, carnosic acid and polydatin against glycation and α-amylase and α-glucosidase enzymes. *Int. J. Food Sci. Technol.* **2018**, *53*, 755-766, https://doi.org/10.1111/ijfs.13579.
48. Gou, S.-H.; Liu, J.; He, M.; Qiang, Y.; Ni, J.-M. Quantification and bio-assay of *α*-glucosidase inhibitors from the roots of *Glycyrrhiza uralensis* Fisch. *Nat. Prod. Res.* **2016**, *30*, 2130-2134, https://doi.org/10.1080/14786419.2015.1114940.
49. Liu, Y.; Zhan, L.; Xu, C.; Jiang, H.; Zhu, C.; Sun, L.; Sun, C.; Li, X. α-Glucosidase inhibitors from Chinese bayberry (Morella rubra Sieb. et Zucc.) fruit: Molecular docking and interaction mechanism of flavonols with different B-ring hydroxylations. *RSC Adv.* **2020**, *10*(49), 29347–29361. https://doi.org/10.1039/d0ra05015f.
50. Yin, Z.; Zhang, W.; Feng, F.; Zhang, Y.; Kang, W. α-Glucosidase inhibitors isolated from medicinal plants. *Food Sci. Hum. Wellness* **2014**, *3*, 136-174, https://doi.org/10.1016/j.fshw.2014.11.003.

51. Zhao, C.; Liu, Y.; Cong, D.; Zhang, H.; Yu, J.; Jiang, Y.; Cui, X.; Sun, J. Screening and determination for potential α-glucosidase inhibitory constituents from *Dalbergia odorifera* T. Chen using ultrafiltration-LC/ESI-MS^n. *Biomed. Chromatogr.* **2013**, *27*, 1621-1629, https://doi.org/10.1002/bmc.2970.
52. Alqahtani, A.S.; Hidayathulla, S.; Rehman, M.T.; ElGamal, A.A.; Al-Massarani, S.; Razmovski-Naumovski, V.; Alqahtani, M.S.; El Dib, R.A.; AlAjmi, M.F. Alpha-Amylase and Alpha-Glucosidase Enzyme Inhibition and Antioxidant Potential of 3-Oxolupenal and Katononic Acid Isolated from *Nuxia oppositifolia*. *Biomolecules* **2020**, *10*, 61, https://doi.org/10.3390/biom10010061.
53. Nguyen, N.-H.; Pham, D.D.; Le, T.-T.-V.; Nguyen, T.-A.-T.; Huynh, D.-L.; Duong, T.-H.; Sichaem, J. Synthesis and α-Glucosidase Inhibitory Activity of Ursolic Acid, Lupeol, and Betulinic Acid Derivatives. *Chem. Nat. Compd.* **2021**, *57*, 1038-1041, https://doi.org/10.1007/s10600-021-03545-1.
54. Priscilla, D.H.; Roy, D.; Suresh, A.; Kumar, V.; Thirumurugan, K. Naringenin inhibits α-glucosidase activity: A promising strategy for the regulation of postprandial hyperglycemia in high fat diet fed streptozotocin induced diabetic rats. *Chem. -Biol. Interact.* **2014**, *210*, 77-85, https://doi.org/10.1016/j.cbi.2013.12.014.
55. Shibano, M.; Kakutani, K.; Taniguchi, M.; Yasuda, M.; Baba, K. Antioxidant constituents in the dayflower (*Commelina communis* L.) and their α-glucosidase-inhibitory activity. *J. Nat. Med.* **2008**, *62*, 349-353, https://doi.org/10.1007/s11418-008-0244-1.
56. Choo, C.Y.; Sulong, N.Y.; Man, F.; Wong, T.W. Vitexin and isovitexin from the Leaves of *Ficus deltoidea* with *in-vivo* α-glucosidase inhibition. *J. Ethnopharmacol.* **2012**, *142*, 776-781, https://doi.org/10.1016/j.jep.2012.05.062.

Publisher's Note & Disclaimer