DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

INTERNATIONAL JOURNAL OF SCIENCE FOR GLOBAL SUSTAINABILITY
(A PUBLICATION OF FACULTY OF SCIENCE, FEDERAL UNIVERSITY GUSAU, NIGERIA)

QSPR Analysis of Infant Infectious Diseases Using Machine Learning

K.O. Aremu[1,2], ‡, S.S. Sarkin Pawa[3], O. Ejima[1], I. Muhammad[1]

[1]Department of Mathematics, Usmanu Danfodiyo University Sokoto, PMB 2346, Sokoto State, Nigeria.

[2]Department of Mathematics and Applied Mathematics, Sefako Makgatho Health Sciences University, P.O. Box 94 Medunsa 0204, Ga-Rankuwa, Pretoria, South Africa.

[3]Department of Computer Science, Federal University, Gusau, Gusau-Zaria Road, P.M.B. 1001, Gusau, Zamfara State, Nigeria.

Corresponding Author's Email: aremukazeemolalekan@gmail.com

Received on: April 2026 **Revised and Accepted on:** June, 2026 **Published on:** July, 2026

ABSTRACT

This study utilizes the neighborhood topological indices to assess the predictive performance of three machine learning models — Linear Regression (LR), k-Nearest Neighbors (KNN), and Random Forest (RF) — to model the relationships between the topological indices and some physicochemical properties (PHPs) of drugs used to treat infant infectious diseases. Root Mean Square Error (RMSE) is used as the evaluation metric across properties such as : boiling point (Bp), molar reactivity (Mr), molar volume (Mv), polarity (P), enthalpy of vaporization (Ev), melting point (Mp), and surface tension (St). The results indicate that kNN consistently provides the best predictive performance, particularly in cases involving complex relationships. LR performs well in linear cases but struggles with non-linear data. RF, although demonstrating strength in isolated cases, tends to exhibit the highest RMSE values. Additionally, the analysis of the Neighborhood Topological Indices (NTIs) shows that NRRG performs better in predicting properties like Bp, Ev, and P, whereas indices such as NFG, NRG, NM1, NSCI, NGA, and NM2 yield more accurate predictions for Mp, St, Mv, Mr, and Fp, respectively. These findings emphasize the importance of selecting appropriate models and indices in QSPR analysis.

Keywords: Infant infectious disease; topological indices; QSPR; Linear Regression; k-Nearest Neighbors; Random Forest

1.0 INTRODUCTION

Infectious diseases remain one of the leading causes of morbidity and mortality among infants worldwide (WHO, 2024). During the first year of life, a child's immune system is still developing, leaving them vulnerable to a wide range of infections. Conditions such as HIV, diarrhea, tuberculosis, measles, mumps, rubella, and chickenpox pose significant health risks to infants due to their immature immune responses and unique physiological susceptibilities. Notably, the immune system of newborns exhibits both quantitative and qualitative differences compared to that of adults, making it more challenging for infants to effectively combat infections. These vulnerabilities highlight the urgent need for improved therapeutic strategies tailored specifically for this age group. Yaser (2024) noted that comprehensive management of infectious disease in infants must encompass prevention, assessment, diagnosis, and treatment, including disease control programs such as vaccination campaigns and antiretroviral therapy for HIV/AIDS. In addition, effective management and control of the transmission of these infant infectious diseases requires information on the incubation period, syndrome, nutrition, and period of infectiousness (Ida et al., 2018; Yehuda & Fernando, 2006; Lawrence, 2018).

The administration of effective drugs is a critical step in mitigating the significant risks posed by infectious diseases in young children. Ensuring drug efficacy not only improves outcomes but also contributes to broader public health goals. However, the development of such efficient drugs is a complex and resource-intensive process, involving substantial financial investment, extensive research, multiple phases of clinical trials, and considerable time to achieve completion (Usman et al., 2024; Aylin et al., 2024; Gamal & Khalid, 2015). These challenges highlight the pressing need for alternatives that are more cost-effective, faster, and less arduous. Structure- and ligand-based drug design approaches provide promising solutions in this regard (Vidushi et al., 2021). These computational methods leverage data derived from molecular structures to predict how small molecules interact with biological targets, enabling the rational design of drug candidates. They incorporate models like Quantitative Structure-Property Relationship (QSPR) and Quantitative Structure-Activity Relationship (QSAR), which utilize statistical and machine-learning techniques to analyze structural descriptors and predict biological activity. By streamlining the drug discovery process, these approaches reduce dependency on traditional experimental methods, offering a more efficient pathway to develop potent drugs (Thereza et al., 2022).

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

Chemical graph theory plays a crucial role in the application of structure-based approaches to drug design and drug property prediction. This branch of graph theory provides a mathematical framework for examining the relationship between graph representations and chemical structures (Vidushi et al., 2021). Molecular graphs serve as visual models of chemical compounds, where vertices represent atoms and edges depict the bonds connecting them (Ernesto & Danail, 2013). A fundamental concept in chemical graph theory is the use of topological indices — numerical descriptors that convert molecular structures into real numbers (Ernesto & Danail, 2013). These indices are widely employed in QSAR and QSPR studies. By leveraging these indices, researchers can predict the biological activities and physical properties of chemical compounds, streamlining the drug discovery process and enhancing the accuracy of molecular property estimation (Ramanathan et al., 2003). This integration of mathematical and computational tools significantly contributes to the efficient design of new drugs with desired biological properties.

Topological indices can be categorized into several types, each offering unique insights into the structural properties of molecules. These include degree-based topological indices, derived from the connectivity of atoms within a molecule; distance-based topological indices, which utilize the shortest paths or distances between pairs of atoms; mixed-based topological indices, which combine features of both degree- and distance-based approaches; and neighborhood degree-based indices, which focus on the degree of adjacent atoms (Abdu & Mohammed, 2021; Mondal et al., 2021). These indices have proven invaluable in predicting the physicochemical properties of drugs used in treating infectious and other diseases, including those targeting mental health, hepatitis, COVID-19, tuberculosis, and Lyme disease, among others (Ejima et al., 2024; Kirana et al., 2024; Khalil et al., 2025; Ugasini et al., 2024; Micheal et al., 2024; Syed et al., 2020; Rongbing et al., 2023; Abubakar et al., 2024).

This research aims to leverage QSPR analysis to examine the physical and chemical properties of drugs used in treating infectious diseases in infants. By employing machine learning models such as Linear Regression (LR), k-Nearest Neighbors (kNN), and Random Forest, the study seeks to predict the pharmacological properties of these drugs based on their topological indices. These predictive models can provide valuable insights into optimizing treatment protocols and improving drug efficacy, which are crucial for enhancing infant health outcomes.

Our research findings provide valuable insights into the interactions between drug compounds, biological systems, and their neighborhood topological indices, with a particular focus on addressing infant infections. The study is structured into four key sections: Preliminaries, which establish the foundational concepts; Materials and Methods, detailing the approaches and tools employed; Main Results, presenting the core findings; and Summary and Conclusion, which synthesize the outcomes and implications of the research.

2.0 PRELIMINARIES

A graph G is composed of a finite, nonempty set V(G) of elements known as vertices, along with a set E(G) of unordered pairs referred to as edges. Topological indices (TIs) are molecular descriptors computed from the corresponding chemical graphs of compounds, with applications in drug design, drug property prediction, and drug efficacy ranking. In this study, we adopt simple, undirected graphs and neighborhood topological indices for analyzing drugs used in treating infant infectious diseases. In the molecular graph representation, atoms are depicted as vertices, and chemical bonds are represented as edges. The neighborhood degree-based topological indices are defined as follows.

Definition 2.1 Let G be a chemical graph with vertices a and b in the vertex set V(G). The neighbor degree sum of a vertex b, denoted by γ_b, is defined as the total sum of the degrees of all vertices adjacent to b. This can be expressed mathematically as:

$$\gamma_b = \sum_{a \in N(b)} d_a \qquad (1)$$

here da is the degree of vertex a, and N(b) represents the set of neighbors of vertex b. The neighborhood degree sum-based topological indices utilized in this study are presented below.

(i) The neighborhood first Zagreb index (Mondal et al., 2019) is given by

$$NM1(G) = \sum_{a,b \in E(G)} (\gamma_a + \gamma_b) \qquad (2)$$

ii) The neighborhood second Zagreb index (Mondal et al., 2019) is given by

$$NM2(G) = \sum_{a,b \in E(G)} (\gamma_a \gamma_b) \qquad (3)$$

iii) The neighborhood forgotten index (Mondal et al., 2019) is given by

$$NF(G) = \sum_{a,b \in E(G)} (\gamma_a^2 + \gamma_b^2) \qquad (4)$$

iv) The neighborhood sum connectivity index (Mondal et al., 2021) is given by

$$NSI(G) = \sum_{a,b \in E(G)} \sqrt{\frac{1}{\gamma_a + \gamma_b}} \qquad (5)$$

v) The neighborhood reciprocal Randić index (Mondal et al., 2021) is given by

$$NRR(G) = \sum_{a,b \in E(G)} \sqrt{\gamma_a \gamma_b} \qquad (6)$$

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

vi) The neighborhood geometric arithmetic index (Abubakar et al., 2022) is given by

$$NGA(G) = \sum_{a,b\in E(G)} \frac{2\sqrt{\gamma_x\gamma_y}}{\gamma_a+\gamma_b} \quad (7)$$

vii) The neighborhood Randić index (Abubakar et al., 2022) is given by

$$NR(G) = \sum_{u,v\in E(G)} \sqrt{\frac{1}{\Phi_u\Phi_v}} \quad (8)$$

viii) The neighborhood atom bond connectivity index (Abubakar et al., 2022) is given by

$$NABC(G) = \sum_{a,b\in E(G)} \sqrt{\frac{\gamma_a+\gamma_b-2}{\gamma_a\gamma_b}} \quad (9)$$

2.1 Linear Regression

Cakra and Trisedya (2015) note that linear regression is a statistical model used to predict the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the variables, meaning the change in the dependent variable is proportional to the change in the independent variables.

1. Dependent Variable (y): The target variable.
2. Independent Variable(s) (X): The predictor.
3. Intercept (β): The constant value where the line intersects the Y-axis when all independent variables are zero.
4. Coefficients (β1, β2, …): The weights that determine the impact of each independent variable on the dependent variable, expressed mathematically as:

$$x = \beta_1\omega_1 + \beta_2\omega_2 + \beta_3\omega_3 + \cdots + \beta_n\omega_n$$

2.1 k-Nearest Neighbors (kNN)

The k-Nearest Neighbors (kNN) technique deals primarily with unlabeled observations by assigning them to the class of the most similar labeled examples. The technique uses various features to characterize both the training and test datasets (Zhongheng, 2016; Cunningham & Delany, 2007). It is a straightforward technique that relies on the identity or properties of the kth neighbors of an unlabeled point for either prediction or classification. There are two important concepts in the kNN algorithm (Zhongheng, 2016): the method of distance calculation and the parameter k. The default method for calculating distance in kNN is Euclidean distance, given by the following equation:

$$D(p,q) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2 + \cdots + (p_n-q_n)^2}$$

where p and q represent different observations in a system. Other distance measures, such as Manhattan distance, are also used. The parameter k defines the number of neighbors considered when making a classification. The choice of k significantly impacts the algorithm's performance; a larger k reduces variance but may ignore important small patterns, while a smaller k is more sensitive to noise. The key is to balance overfitting and underfitting, and some suggest setting k to the square root of the number of observations in the training dataset.

2.1.1 Random Forest

Random forest serves as a potent ensemble learning technique consisting of multiple decision trees, renowned for its robustness and accuracy in classification tasks. In contrast to a single decision tree, which can be prone to overfitting or biased towards specific features, random forest combines the predictions of numerous trees to yield a more dependable result. In a random forest, each tree autonomously casts a vote on the class for a given input vector (x). The final prediction of the forest is then determined by the majority vote across all the trees. Users can define the number of trees (N) in the forest, allowing for flexibility and optimization based on the specific problem (Gök & Olgun, 2021):

$$R_f = majority\ vote(c_b(x))$$

3.0 MATERIALS AND METHODS

In this study, eight neighborhood degree-based topological indices (NTIs), namely the neighborhood first Zagreb index (NM1(G)), the neighborhood second Zagreb index (NM2(G)), the neighborhood forgotten index (NF(G)), the neighborhood sum connectivity index (NSCI(G)), the neighborhood reciprocal Randić index (NRR(G)), the neighborhood geometric arithmetic index (NGA(G)), the neighborhood Randić index (NR(G)) and the neighborhood version of the atom bond connectivity index (NABC(G)), were computed and employed to evaluate the predictive performance of three machine learning models: Linear Regression (LR), k-Nearest Neighbors (KNN) and Random Forest (RF). These models were applied to Quantitative Structure-Property Relationship (QSPR) analysis, focusing on eight physical properties of seventeen (17) drugs used in the treatment of infant infectious diseases. These properties are boiling point (Bp), melting point (Mp), flash point (Fp), enthalpy of vaporization (Ev), molar refraction (Mr), polarization (P), surface tension (St), and molar volume (Mv). The algorithm below provides an overview of the steps adopted in this study.

Algorithm 1: Overview of Models Implementation.

1. Load dataset {X, y} from neighborhood degree sum topological indices and physical properties of drugs used in the treatment of infectious diseases.
2. Scale data to control outliers.
3. Split dataset into training and testing sets using train-test split: {X_train, X_test, y_train, y_test} ← train_test_split(X, y, test_size = 0.3).
4. Model building — initialize hyperparameter grid for KNN: param_grid_KNN ← {'k' = 1}.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

5. Model building — initialize hyperparameter grid for LR: param_grid_LR ← default parameters.
6. Model building — initialize hyperparameter grid for RF: param_grid_RF ← {'bootstrap': [True], 'n_estimators': 50, 'max_features': [sqrt], 'max_depth': [none], 'max_leaf_nodes': none, 'min_samples_leaf': 4, 'min_samples_split': 10, 'learning_rate': 0.1}.
7. Add model_KNN, model_LR, and model_RF to models.
8. Initialize base models: base_model_regressor ← KNN, LR and RF(models).
9. Fit base models: base_model_regressor.fit(X_train, y_train).
10. Predict physical properties for testing set: ŷ_test ← base_model_regressor.predict(X_test).
11. Calculate model performance metrics (Root Mean Squared Error) for ŷ_test and y_test.

3.1 Molecular Structures of Some Drugs Used in Treating Infant Infectious Diseases

This study examines sixteen drugs commonly used in the treatment and management of infant infectious diseases. Figure 1 presents the illustrations of three of these drugs.

infectious diseases. However, the following Figure ?? presents the illustrations of three of these drugs.

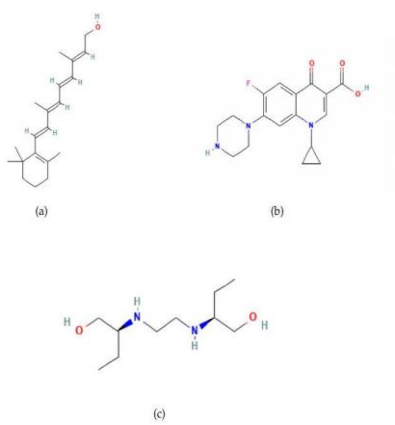

Figure 1: Molecular structure of drug (a) Vitamin A (b) Ciprofloxacin (c) Ethambulton.

4. Main Results

. Figure 1: Molecular structure of drug (a) Vitamin A (b) Ciprofloxacin (c) Ethambutol

4.0 RESULTS

In this section, we present the research results. The physical property dataset was obtained from ChemSpider (www.chemspider.com), a comprehensive database of chemical information (see Table 1). Molecular structures of the selected drugs were used to calculate the neighborhood degree-based topological indices (see Table 3). The data in both Tables 1 and 3 form the foundation for the QSPR models.

Table 1: Drugs and Physical Properties

Drugs	BP	MP	EVP	PL	MV	MR	SF	FP
Vitamin A	421.2	63.5	78.0	37.9	36.9	300.0	95.5	147.3
Paracetamol	387.8	169.0	66.2	16.8	52.8	120.9	42.4	188.4
Ciprofloxacin	581.8	256.0	91.5	33.0	67.4	226.8	83.3	305.6
Isoniazid	251.97	0.0	1720.0	14.6	57.8	110.0	36.9	190.0
ORS	405.4	158.0	69.3	16.2	60.5	117.9	41.0	199.0
Amoxicillin	333.0	129.0	57.6	20.3	44.3	148.4	51.2	155.2
Hydrocortisone	743.2	194.0	113.7	36.3	85.3	236.2	91.5	403.3
Ethionamide	566.0	220.0	97.7	37.9	58.8	281.4	95.6	310.4
Levofloxacin	571.5	224.0	90.1	36.1	70.3	244.0	91.1	299.4
Linezolid	585.0	177.0	87.5	32.9	47.7	259.0	83.5	307.9
Moxifloxacin	636.0	270.0	98.8	40.4	60.6	285.0	101.8	338.7
Terizidone	569.2	175.0	92.1	30.2	62.5	198.9	76.1	298.1
FLAGYL	405	158	69.3	16.2	60.5	117.9	41	99
Ethambutol	188	202	63.4	39.5	72.3	154.4	59.3	180
Abcavir	300	165	81.4	76.2	173.6	61.39	nan	
Aciclovir	300	256	73.4	34.5	74.2	151.4	55.11	nan
Lamivudine	320	180	84.5	Nan	78.5	158.6	56.49	nan

In the following section, we provide a detailed illustration of a chemical graph using the molecular graph representation of Vitamin A. This serves as an example to visualize the structure of chemical graphs. Following this, we demonstrate the step-by-step computation of neighborhood topological indices, utilizing the formulas outlined in Definition 2.1.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

Linezolid	585.0	177.0	87.5	32.9	47.7	259.0	83.5	307.9
Moxifloxacin	636.0	270.0	98.8	40.4	60.6	285.0	101.8	338.7
Terizidone	569.2	175.0	92.1	30.2	62.5	198.9	76.1	298.1
FLAGYL	405	158	69.3	16.2	60.5	117.9	41	99
Ethambul	188	202	63.4	39.5	72.3	154.4	59.3	180
Abcavir	300	165	81.4	76.2	173.6	61.39	nan	
Aciclovir	300	256	73.4	34.5	74.2	151.4	55.11	nan
iamividun	320	180	84.5	nan	78.5 158.6	56.49	nan	nan

ıe following section, we provide a detailed illustration of a chemical graph using the molecu
ɔh representation of Vitamin A. This serves as an example to visualize the structure of chemi
ɔhs. Following this, we demonstrate the step-by-step computation of neighborhood topologi
ces, utilizing the formulas outlined in Definition 2.1.

Figure 2: Chemical graph of Vitamin A.

The following table presents the edge partitioning of the chemical graph of Vitamin A presented in Figure 2.

$$|V| = \{v1, v2, v3, v4, v5, v6, v7, v8, v9, v10, v11, v12, v13, v14, v15, v16, v17, v18, v19, v20, v21\}$$

Table 2: Vertex and Edge Partitioning Data

Edge	Label	(du, dv)	m
v1v3	a	(4,7)	2
v3v4	b	(7,9)	1
v4v5	c	(9,6)	1
v5v9	d	(6,3)	1
v5v6	e	(6,5)	1
v7v6	f	(5,4)	1
v7v8	g	(5,5)	5
v8v3	h	(5,3)	3
v4v10	i	(3,2)	1
v10v11	j	(5,4)	2
v12v13	k	(4,6)	1
v18v19	l	(6,7)	1
v19v20	m	(9,5)	1

Theorem 4.1 Let G be the chemical graph of Vitamin A such that u and v are elements of the vertex set V(G) of G. Then:

1. NM1(G) = 216
2. NM2(G) = 568
3. NF(G) = 1214
4. NSCI(G) = 6.695
5. NRR(G) = 106.16
6. NR(G) = 4.419
7. NGA(G) = 20.63
8. NABC(G) = 12.080

Proof. Adopting the formulas in Definition 2.1 and the edge partitioning data in Table 2, we have the following:

$$\sum_{a,b\in E(G)} (\gamma_a + \gamma_b)\ 22 + 16 + 15 + 9 + 11 + 9 + 50 + 24 + 5 + 18 + 10 + 13 + 14 = 216$$

_a,b∈E(G)?γ_aγ_b= 56+63+54+18+30+125+45+6+40+24+42+45+20 = 568

$$\sum_{a,b\in E(G)} (\gamma_a^2 + \gamma_b^2)\ 130 + 130 + 117 + 45 + 61 + 41 + 250 + 102 + 13 + 82 + 52 + 85 + 106 = 1214$$

_a,b∈E(G)?&1/γ_a+γ_b= 6.695

$$\sum_{a,b\in E(G)} \sqrt{\gamma_a\gamma_b}\ \sqrt{28} + \sqrt{63} + \sqrt{54} + \sqrt{18} + \sqrt{30} + \sqrt{20} + \sqrt{25} + \sqrt{15} + \sqrt{6} + \sqrt{20} + \sqrt{24} + \sqrt{42} + \sqrt{45} = 106.16$$

_a,b∈E(G)?&1/γ_aγ_b= 4.419

$$\sum_{a,b\in E(G)} \frac{2\sqrt{\gamma_a\gamma_b}}{\gamma_a + \gamma_b}\ 20.63$$

_a,b∈E(G)?&γ_a+γ_b−2/γ_aγ_b= 12.080

These steps in the proof of Theorem 4.1 were followed to obtain the NTIs of the remaining drugs as contained in Table 3.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

Table 3: Drugs and Computed Topological Indices

Drug	NM1(G)	NM2(G)	NFG(G)	NGA(G)	NSI(G)	NABC(G)	NR(G)	NRR(G)
Vitamin A	216	568	1214	20.6	6.69	12.08	106.1	4.41
Paracetamol	106	259	530	10.91	3.58	6.386	52.65	2.37
Ciprofloxacin	328	1022	2132	26.6	7.9	14.53	162	4.76
Isoniazid	96	234	496	9.82	3.28	5.89	47.22	2.23
ORS	128	356	766	11.731	3.8	6.88	62.69	2.54
Amoxicillin	336	1076	2278	26.4	7.8	14.5	165.16	4.7
Hydrocortisone	396	1379	3006	28.21	8.07	15.18	193.12	4.74
Ethionamide	110	280	588	10.81	3.54	6.388	54.20	2.362
Levofloxacin	360	1152	2412	28.54	8.41	15.53	177.55	5.04
Linezolid	288	819	1698	25.73	7.95	14.42	142.61	4.98
Moxifloxacin	414	1340	2796	32.5	9.54	17.59	204.37	5.72
Terizidone	256	680	1408	23.76	7.38	13.41	126.81	4.61
Flagyl	128	356	766	11.7	3.8	6.8	62.88	2.544
Lamivudine	176	493	1028	15.76	4.9	8.9	86.93	3.10
Abcavir	286	873	1794	23.77	7.07	12.92	141.88	4.27
Aciclovir	186	534	1112	16.78	5.30	9.54	91.94	3.43

To implement a QSPR model for predicting the physical properties of drugs used in treating infant infectious diseases, we utilized three regression algorithms: Linear Regression (LR), k-Nearest Neighbors (k-NN) and Random Forest (RF). The objective was to develop a predictive model using topological indices as the input predictors and various physical properties, such as boiling point (BP) and melting point (MP), as the target responses. These regression models were trained and evaluated to assess their effectiveness in predicting these critical drug properties based on the specified indices. To begin the implementation, scatter plots of the dataset were generated to determine the extent of the relationship between the computed neighborhood topological indices and the physical properties of these drugs. Figure 3 below shows some of these correlations observed.

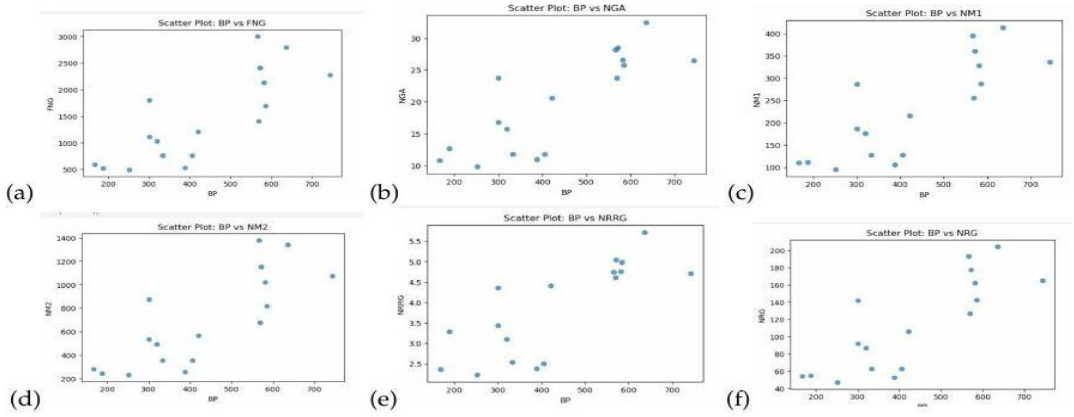

Figure 3: Correlation plots of (a) NFG(G) vs Boiling Point, (b) NGA(G) vs Boiling Point, (c) NM1(G) vs Boiling Point, (d) NM2(G) vs Boiling Point, (e) NRG(G) vs Boiling Point and (f) NRR(G) vs Boiling Point

The presented scatter plots illustrate the relationships between the drugs' physical properties and their molecular descriptors. Specifically, the boiling point (BP) is used to demonstrate this relationship. Across all plots, the data points reflect varying degrees of association, with some showing more defined patterns than others. In general, there is evidence of both linear and nonlinear trends. These observations of different trends support the choice of models (Linear Regression, k-Nearest Neighbors and Random Forest) to ensure an in-depth analysis of the data.

4.1 Models Implementation

In this subsection, we present the implementation of quantitative structure-property relationship (QSPR) using three machine learning models: Linear Regression (LR), k-Nearest Neighbors (KNN), and Random Forest (RF). The process begins with the loading of a dataset containing topological indices derived from molecular structures and the physical properties of drugs used for treating infant infectious diseases. These indices and properties form the input-output pairs, X and y, for model

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

training and testing.

Before model training, the data is scaled to control outliers and ensure uniformity across features. The dataset is then split into training and testing sets using a train-test split method, with 70% assigned to training and 30% to testing.

Subsequently, to ensure optimal performance of the models such that the risk of either over-fitting or under-fitting is highly reduced, the best hyperparameters are selected using the GridSearch technique. First, for the K-Nearest Neighbors (KNN) Regressor, the following parameters were considered: algorithm ∈ {auto, ball_tree, kd_tree, brute}; metric ∈ {euclidean, manhattan, minkowski}; k ∈ range(1, 5, 2); weight ∈ {uniform, distance}. The best hyperparameters obtained for the KNN are: Algorithm — auto (automatically selects the best algorithm based on data size); Metric — euclidean (uses Euclidean distance for nearest neighbor calculation); k = 1 (considers only the closest neighbor); Weights — uniform (all points in each neighborhood are weighted equally).

Similarly, for the RF regressor, the following options were considered: bootstrap ∈ {True}; n_estimators ∈ range(10, 100, 10); max_depth ∈ {none}; min_samples_split ∈ range(1, 10, 2); learning_rate ∈ range(0.05, 0.1, 20); max_features ∈ {sqrt, log2, None}; min_samples_leaf ∈ range(2, 10, 2). The search returned the following as the best hyperparameters for the RF: n_estimators = 50, max_depth = none, min_samples_split = 10, learning_rate = 0.1, max_features = sqrt, min_samples_leaf = 4. However, no hyperparameters were tuned for the Linear Regression, as it was implemented based on a closed-form solution to minimize errors.

Furthermore, these models were then trained on the training data (X_train, y_train) to establish their respective predictive relationships. Afterwards, the models were used to predict the physical properties of the testing dataset (y_test) using the corresponding feature inputs (X_test).

Finally, the performance of each model is evaluated using the Root Mean Squared Error (RMSE) metric, which measures the deviation of predicted values from the actual values in the test set. The QSPR analysis provides insights into how well each model captures the relationship between the molecular descriptors and the physical properties, allowing for comparison of their predictive accuracies.

5.0 PERFORMANCE METRIC ANALYSIS

Below we present tables showing the performance of each of the NTI and the various regression models considered in this study — Root Mean Square Error (RMSE) across models and properties.

NM1

PHP	LR	KNN	RF
Bp	133.22	127.72	156.40
Mp	65.17	62.23	64.09
Ev	10.27	10.39	10.69
P	5.39	6.03	8.85
St	16.41	15.54	18.37
Mv	51.45	60.49	60.24
Mr	14.36	17.69	17.29
Fp	61.05	60.90	65.12

NM2

PHP	LR	KNN	RF
Bp	138.39	140.21	160.47
Mp	62.50	61.93	61.60
Ev	10.88	11.29	10.53
P	5.64	6.69	9.89
St	16.43	15.93	18.84
Mv	54.92	61.98	69.22
Mr	15.80	18.99	20.50
Fp	61.37	61.68	56.95

NFG

PHP	LR	KNN	RF
Bp	137.56	140.21	156.64
Mp	62.52	61.93	60.60
Ev	10.92	11.29	10.38
P	5.60	6.69	8.25
St	16.45	15.93	18.24
Mv	54.10	61.98	68.77
Mr	15.53	18.99	20.00
Fp	62.01	61.68	56.99

NGA

PHP	LR	KNN	RF
Bp	131.41	130.19	148.38
Mp	68.81	76.04	65.29
Ev	10.12	10.91	10.92
P	5.15	4.97	6.63
St	16.38	16.57	19.18
Mv	47.70	40.83	59.01
Mr	12.95	10.69	16.18
Fp	63.66	80.42	70.98

NSCI

PHP	LR	KNN	RF
Bp	132.07	129.16	162.23
Mp	70.91	77.81	89.22
Ev	10.34	10.89	12.14
P	4.90	5.28	6.56
St	16.35	16.56	15.60
Mv	44.34	45.49	40.49
Mr	11.82	12.07	11.93
Fp	67.08	80.60	82.95

NABC

PHP	LR	KNN	RF
Bp	130.49	129.16	161.44
Mp	70.42	77.81	69.48
Ev	10.17	10.89	11.50

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

PHP	LR	KNN	RF
P	4.93	5.28	6.48
St	16.36	16.56	15.56
Mv	44.69	45.49	44.19
Mr	11.86	12.07	12.00
Fp	65.96	80.60	83.16

NRG

PHP	LR	KNN	RF
Bp	133.64	127.72	158.56
Mp	65.19	62.23	64.20
Ev	10.28	10.39	10.74
P	5.42	6.03	8.86
St	16.40	15.54	18.48
Mv	51.76	60.69	60.99
Mr	14.47	17.69	17.54
Fp	60.94	60.90	65.45

NRRG

PHP	LR	KNN	RF
Bp	137.40	114.43	160.16
Mp	73.39	79.72	81.38
Ev	10.91	9.82	12.24
P	4.59	4.92	6.45
St	16.34	16.48	16.25
Mv	40.96	45.55	49.13
Mr	10.92	12.09	14.09
Fp	72.26	75.05	84.88

The analysis of the RMSE values for the three models — Linear Regression (LR), K-Nearest Neighbors (KNN), and Random Forest (RF) — across various physical properties (PHPs) of drugs used in the treatment of infant infectious diseases shows that model performance varies significantly depending on the PHPs and the topological indices employed.

The Linear Regression (LR) model exhibits stable performance with moderate RMSE values across the NTIs. It tends to perform well in simpler cases but struggles with more complex patterns. k-Nearest Neighbors (KNN), on the other hand, often outperforms LR in capturing relationships with better RMSE values, while Random Forest (RF) tends to have the highest RMSE in most cases. Specifically, in NM1 and NM2, KNN and LR consistently outperform RF, although RF recorded the lowest RMSE of 56.95 for Fp in NM2, suggesting its strength in some isolated cases. In NFG and NGA, LR and KNN provide competitive performance, with KNN slightly ahead in several properties, while RF shows higher RMSE values, indicating weaker predictive power for these NTIs.

In NRRG, kNN presents some interesting results, achieving the lowest RMSE for Bp (114.43) and Ev (9.82), while LR performs better in the prediction of P (4.59), demonstrating its strength in handling linearly related data. For NSCI and NRG, LR and KNN have similar RMSE values, but RF exhibits significantly higher RMSE in most cases, although it performs better in some instances — notably achieving the lowest RMSE for Mv in NSCI (40.49), confirming its effectiveness in select cases.

On the other hand, analysis of the NTIs reveals that NRRG demonstrates superior performance in predicting properties such as Bp, Ev, and P. Meanwhile, indices like NFG, NRG, NM1, NSCI, NGA, and NM2 show greater accuracy in predicting Mp, St, Mv, Mr, and Fp, respectively.

Overall, kNN appears to be the most consistently well-performing model across all the NTIs, particularly in cases where complex relationships exist. Linear Regression provides a reliable baseline but struggles in non-linear cases. Random Forest, while sometimes effective for specific properties, tends to have the highest RMSE in most cases, suggesting relative inefficiency for this study's dataset.

In the next section we present predictions for the 30% test data of this study.

Table 4: Predicted and Actual Properties of Various Drugs

PHP	Vit A (Pred.)	Vit A (Actual)	Paracetamol (Pred.)	Paracetamol (Actual)	Amoxicillin (Pred.)	Amoxicillin (Actual)	Aciclovir (Pred.)	Aciclovir (Actual)	Terizidone (Pred.)	Terizidone (Actual)
Bp	498.46	421.20	295.40	387.80	588.06	743.00	282.60	300.00	588.06	569.20
Mp	167.24	63.50	186.05	169.00	236.11	194.00	167.09	256.00	174.37	175.00
Ev	86.04	78.00	69.41	66.20	93.12	113.70	65.84	73.40	93.12	92.10
P	33.86	37.90	23.91	16.80	35.34	36.30	29.08	34.50	34.84	30.20
St	60.66	36.90	60.66	52.80	60.96	85.30	64.80	74.20	64.88	62.50
Mv	236.70	300.00	141.33	120.90	236.70	236.20	152.17	151.40	236.70	198.90
Mr	82.00	95.50	48.30	42.40	91.06	91.50	52.22	55.11	82.00	76.10
Fp	200.16	147.30	170.17	188.40	304.91	403.30	198.65	241.45	267.75	298.10

In what follows are the graphs of the predicted values versus the actual values of the physical properties of the drugs, presented to show the pictorial relationship between these values.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

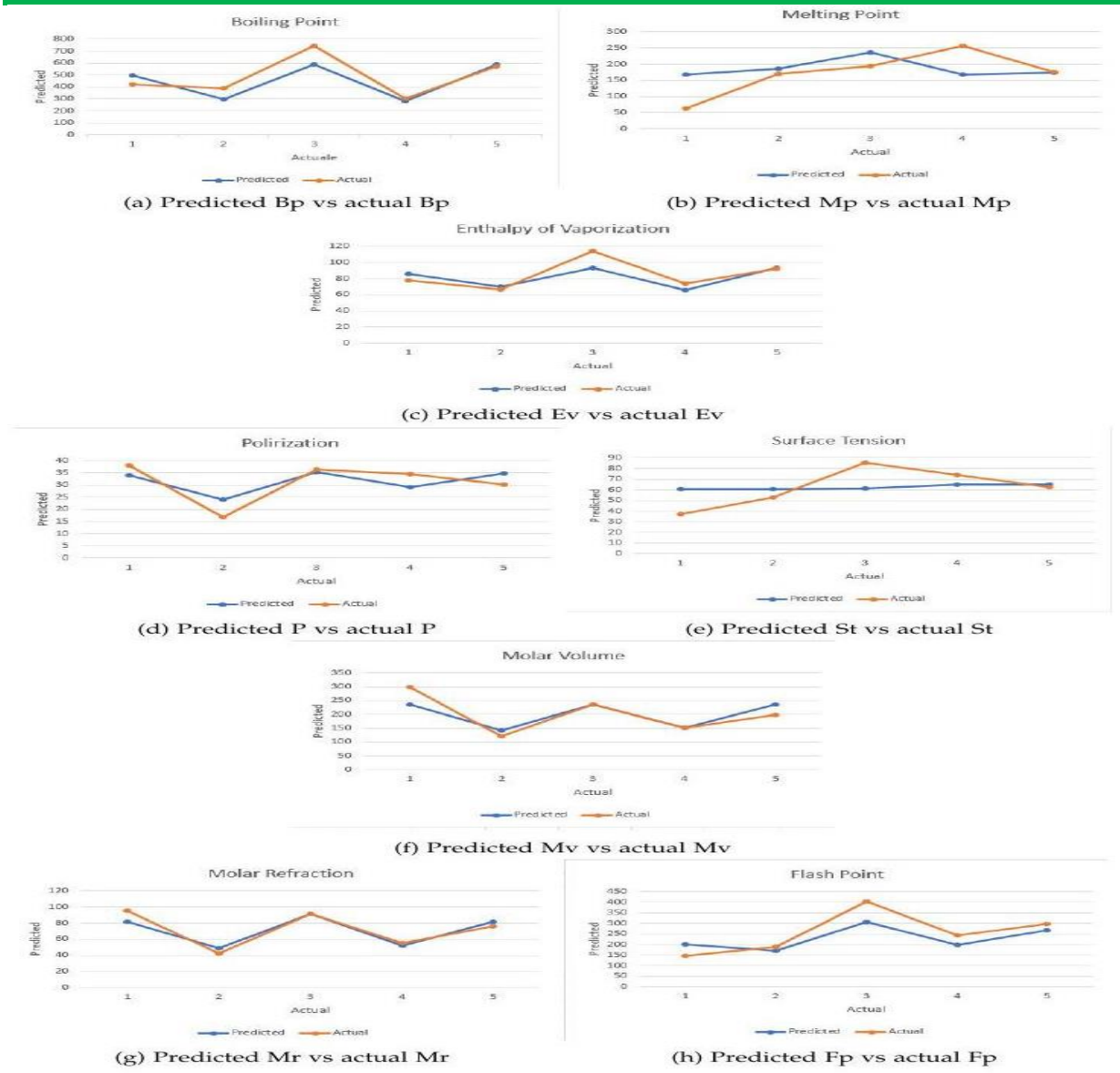

Figure 4: Predicted vs Actual — (a) Boiling Point, (b) Melting Point, (c) Enthalpy of Vaporization, (d) Polarization, (e) Surface Tension, (f) Molar Volume, (g) Molar Refraction, (h) Flash Point.

6.0 SUMMARY AND CONCLUSION

This study applies Quantitative Structure Property Relationship (QSPR) analysis to examine the physical and chemical properties of 17 drugs used to treat infectious diseases in infants, employing three machine learning models: Linear Regression (LR), k-Nearest Neighbors (kNN), and Random Forest (RF). The research aims to predict physicochemical properties such as boiling point, molar volume, enthalpy of vaporization, and flash point, among others, based on topological indices.

The results reveal that kNN consistently outperforms the other models in capturing complex relationships, while LR provides stable performance but struggles with non-linearity. RF, although effective in isolated cases, generally exhibits the highest RMSE values, indicating weaker predictive power across most topological indices. Overall, kNN emerges as the most reliable model for predicting these drug properties, particularly in cases involving complex relationships. LR serves as a strong baseline for linear cases, whereas RF, despite its occasional strengths, demonstrates higher prediction errors in most instances. These findings emphasize the

significance of selecting suitable predictive models for different drug properties to enhance drug efficacy and optimize treatment strategies.

ACKNOWLEDGMENT
The authors appreciate the valuable suggestions of the reviewers in improving the manuscript.

DISCLOSURE OF FUNDING
This research was funded by Usmanu Danfodiyo University Sokoto Institution-Based Research (UDUS-IBR) grant by the Tertiary Education Trust Fund (TETFUND).

CONFLICT OF INTEREST
The authors affirm that there are no conflicts of interest.

DATA AVAILABILITY
No data from this study was deposited into a publicly accessible repository. The data supporting the findings of this study are provided within this manuscript.

REFERENCES
Abdu, A., & Mohammed, A. (2021). Topological indices types in graphs and their applications. Generis Publishing.

Abubakar, M. S., Aremu, K. O., & Aphane, M. (2022). Neighborhood versions of geometric–arithmetic and atom bond connectivity indices of some popular graphs and their properties. Axioms, 11(9), 487.

Abubakar, M. S., Aremu, K. O., Aphane, M., & Amusa, L. B. (2024). A QSPR analysis of physical properties of antituberculosis drugs using neighbourhood degree-based topological indices and support vector regression. Heliyon, 10(7), e28260.

Aylin, S., Trinidad, B., Amber, J., & Benjamin, D. S. (2024). Costs of drug development and research and development intensity in the US, 2000–2018. Statistics and Research Methods, 7(6), e2415445.

Cakra, Y. E., & Trisedya, B. D. (2015). Stock price prediction using linear regression based on sentiment analysis. In 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS) (pp. 147–154). IEEE.

Cunningham, S. J., & Delany, S. J. (2007). k-Nearest neighbour classifiers. ACM Computing Surveys, 54(6).

Ejima, O., Abubakar, M. S., Sarkin Pawa, S., Ibrahim, A., & Aremu, K. O. (2024). Ensemble learning and graph topological indices for predicting physical properties of mental disorder drugs. Physica Scripta.

Ernesto, E., & Danail, B. (2013). Chemical graph theory. In Handbook of graph theory.

Gamal, O. E., & Khalid, O. A. (2015). Drug development: Stages of drug development. Pharmacovigilance, 3(3), e160.

Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of COVID-19 patients with blood samples. Neural Computing and Applications, 33(22), 15693–15707.

Ida, C., Chantal, Q., Pierluigi, L., Jan, C. S., & ECDC Expert Panel Working Group. (2018). Management and control of communicable diseases in schools and other child care settings: Systematic review on the incubation period and period of infectiousness. BMC Infectious Diseases, 18, 199.

Khalil, H. H., Abdul, R. K., & Muhammad, A. (2025). Mathematical modeling and QSPR analysis of hepatitis treatment drugs through connection indices: An innovative approach. Heliyon, 11(1), e41234.

Kirana, B., Shanmukha, M. C., & Usha, A. (2024). A QSPR analysis and curvilinear regression models for various degree-based topological indices: Quinolone antibiotics. Heliyon, 10(12), e32397.

Lawrence, D. F. (2018). Infectious diseases as a cause of global childhood mortality and morbidity: Progress in recognition, prevention, and treatment. Advances in Pediatric Research, 5, 4.

Micheal, A., Joseph, C. H., Berin, G. A., Muhammad, U. G., Gajavalli, S., Fairouz, T., et al. (2024). QSPR analysis of distance-based structural indices for drug compounds in tuberculosis treatment. Heliyon, 10(2), e23981.

Mondal, S., De, N., & Pal, A. (2019). On some general neighborhood degree based topological indices. International Journal of Applied Mathematics, 32(6), 1037.

Mondal, S., Dey, A., De, N., & Pal, A. (2021). QSPR analysis of some novel neighbourhood degree-based topological descriptors. Complex & Intelligent Systems, 7, 977–996.

Ramanathan, N., Kamalakanan, P., & Nirdosh, I. (2003). Applications of topological indices to structure-activity relationship modelling and selection of mineral collectors. Indian Journal of Chemistry, 42(6), 1330–1346.

Rongbing, H., Abid, M., Muhammad, W. R., Sajid, M. A., & Muhammad, K. S. (2023). On molecular modeling and QSPR analysis of lyme disease medicines via topological indices. The European Physical Journal Plus, 134, 243.

Syed, A. K., Kirmani, P. A., & Faizul, A. (2020). Topological indices and QSPR/QSAR analysis of some antiviral drugs being investigated for the treatment of COVID-19 patients. International Journal of Quantum Chemistry, 121(9), e26594.

DOI: https://doi.org/10.57233/ijsgs.v12i2.1123 ISSNp: 2488-9229; ISSNe: 3027-1118

IJSGS FUGUSAU VOL 12 (2) WEBSITE: https://fugus-ijsgs.com.ng

Thereza, A. S., Ariane, N. A., Mazzolari, A., Fiorella, R., & Wei, G.-W. K. (2022). The (re)-evolution of quantitative structure–activity relationship (QSAR) studies propelled by the surge of machine learning methods. Journal of Chemical Information and Modeling, 62, 5317–5320.

Ugasini, P. P., Suresh, M., Fikadu, T. T., & Ebenezer, B. (2024). QSPR/QSAR study of antiviral drugs modeled as multigraphs by using TI's and MLR method to treat COVID-19 disease. Scientific Reports, 14, 13150.

Usman, S., Aisha, A., Madiha, A., Nosheen, A., Mohammed, S. A., Soad, K. A. J., et al. (2024). A comprehensive review of discovery and development of drugs discovered from 2020–2022. Saudi Pharmaceutical Journal, 32(1), 101913.

Vidushi, S., Sharad, W., & Hirdesh, K. (2021). Structure- and ligand-based drug design: Concepts, approaches, and challenges. In Chemoinformatics and bioinformatics in the pharmaceutical sciences. Elsevier.

World Health Organization. (2024). Communicable diseases among children. Retrieved December 2024, from https://www.who.int/teams/maternal-newborn-child-adolescent-health-and-ageing/child-health/communicable-diseases-among-children

Yaser, M. A. (2024). Infectious diseases: Overview. In Handbook of medical and health sciences in developing countries (pp. 1–23).

Yehuda, B., & Fernando, S. (2006). Integrated management of childhood illness: An emphasis on the management of infectious diseases. Seminars in Pediatric Infectious Diseases, 17, 80–98.

Zhongheng, Z. (2016). Introduction to machine learning: k-nearest neighbors. Annals of Translational Medicine, 4(11), 218.