

Machine Learning-Based Analysis of Factors Influencing Oral Health

Kevin Heo^{1,2*}, Lenny Park³ and Joonghyeok H⁴

¹STEM Academy, 4940 South Rd, Odessa, Texas 79762, USA

²Department of Biology, University of Texas Permian Basin, Odessa, Texas, 79762, USA

³North Penn High School, 1340 S Valley Forge Rd, Lansdale, PA 19446, USA

⁴Department of Geoscience, University of Texas Permian Basin, Odessa, Texas, 79762, USA

***Corresponding author:** : Kevin Heo, STEM Academy, 4940 South Rd, Odessa, Texas 79762, Department of Biology, University of Texas Permian Basin, Odessa, Texas, 79762, USA

ARTICLE INFO

Received:  October 13, 2026

Published:  October 28, 2026

Citation: Kevin Heo, Lenny Park and Joonghyeok H. Machine Learning-Based Analysis of Factors Influencing Oral Health. Biomed J Sci & Tech Res 67(2)-2026. BJSTR.MS.ID.010460.

ABSTRACT

This study aimed to apply an exploratory machine learning framework to identify chemical and geochemical factors that may influence oral and dental health. Long-term historical water-quality records comprising 505 samples, collected between March 1940 and August 2026, were compiled and analyzed for parameters using a Python-based machine learning workflow implemented in Google Colab, including descriptive statistics, boxplot and frequency-distribution analysis, and correlation-matrix modeling. Fluoride, nitrate, pH, and total dissolved solids showed the strongest relevance to oral health, each displaying right-skewed distributions and frequent deviations beyond recommended guideline thresholds. Correlation analysis revealed a strong geochemical association between fluoride and total alkalinity, sodium, and total dissolved solids, suggesting shared mineralization pathways rather than a common anthropogenic source. These patterns were broadly consistent with, yet more heterogeneous than, findings reported in prior fluoride- and nitrate-focused groundwater studies, reflecting the extended temporal and spatial coverage of the dataset. These findings demonstrate that machine learning-assisted screening of routinely collected water-quality data can efficiently flag chemical indicators of oral health relevance, offering a scalable, low-cost complement to conventional epidemiological surveillance and a foundation for future predictive, individualized oral health risk modeling.

Keywords: Artificial Intelligence; Machine Learning; Water Quality; Dental Health; Oral Health; Texas

Introduction

Oral health is a fundamental component of overall well-being, influencing nutrition, speech, self-esteem, and systemic health outcomes [1]. According to the World Health Organization [2], oral diseases affect nearly 3.5 billion people worldwide, with untreated dental caries remaining the single most prevalent health condition globally. Beyond its direct burden, poor oral health has increasingly been linked to systemic conditions such as cardiovascular disease, diabetes, and adverse pregnancy outcomes, underscoring the need for a broader, exposure-based understanding of the environmental determinants of oral disease. Machine learning has increasingly been applied to environmental water-quality assessment, offering scalable alternatives to costly, time-intensive laboratory-based monitoring [3-

5]. Fluoride- and nitrate- focused studies employing random forest, support vector machine, and ensemble algorithms have successfully predicted groundwater contamination risk and identified key geochemical drivers [6-8]. In parallel, machine learning has been applied directly to oral health outcomes, including dental caries risk prediction using demographic, behavioral, and clinical features [9-12].

However, very few studies have bridged these two research streams to jointly examine machine learning-based analysis of water chemistry parameters in relation to oral health risk indicators. However, there are problems in these studies: prior groundwater machine learning research has largely focused on single contaminants, such as fluoride or nitrate in isolation, evaluated against generic drinking-water safety thresholds, rather than jointly across a broad, multi-param-

eter chemical profile relevant to oral health. Previous studies have a limited attribution examining the impacts of temporally extensive and spatially diverse groundwater datasets, most having been restricted to short monitoring windows or narrow geographic regions. No detection studies have been undertaken with a parsimonious, exploratory machine learning framework designed specifically to screen and rank multiple chemical parameters by their likely relevance to oral and dental health, using accessible tools such as Google Colab that support reproducibility and public health translation.

This study contributes the application of an exploratory, reproducible machine learning workflow to an unusually long-term, 86-year, and spatially extensive dataset and the joint characterization of 15 parameters relevant to oral health. It also provides an interpretable statistical and visual foundation, via machine learning-based analysis, that can guide future predictive modeling of oral health risk. The objectives of this study are:

1. To characterize the statistical distribution and variability of 15 physicochemical parameters relevant to oral health using descriptive and machine learning-assisted visualization techniques;
2. To identify the parameters most strongly associated with oral, dental, and human health based on their distributional characteristics and known toxicological or dental relevance; and
3. To examine the correlation structure among these parameters to elucidate shared geochemical pathways that may compound or modify oral health risk.

Data and Methods

Data were obtained from the Texas Water Development Board (TWDB) Groundwater Database, a publicly available, long-term repository of water-quality observations from selected wells, springs, and oil/gas tests across the state [13]. Historical records were collected for 15 physicochemical parameters across 505 samples drawn from 238 unique well identifiers, spanning March 1940 to August 2026. The parameters were Calcium, Magnesium, Sodium, Potassium, TDS (log₁₀), Total Alkalinity, Total Hardness (log₁₀), Carbonate, Temperature, Fluoride, Nitrate, Well Depth, Chloride, Sulphate, and pH. Machine-learning and statistical analyses were implemented in Google Colab using Python. The analytical workflow comprised data cleaning and log-transformation of skewed variables (TDS, Total Hardness), computation of descriptive statistics (mean, standard deviation, minimum, maximum), normality testing, and construction of the visual figures presented in this study: a box-and-whisker plot, a frequency-distribution panel, and a Pearson correlation heatmap. Parameter significance was assessed against recognised drinking-water guidelines from the WHO and the U.S. Environmental Protection Agency [14-16]. Table 1 summarises the descriptive statistics for the 15 parameters. Each parameter was tested for statistical significance of its distribu-

tional deviation from a reference baseline; the reported p-value indicates whether the observed central tendency and variance were unlikely to have arisen by chance under the null hypothesis of no systematic structure. A p-value below 0.05 denotes a highly significant parameter—that is, a variable whose distribution carries meaningful, non-random information—whereas the non-significant result for Total Hardness (p = 0.14) indicates greater distributional homogeneity and weaker evidence of structured variation across wells.

Table 1: Summary of 15 Parameters.

Parameters	Unit	Mean	Std	Min	Max	p-value
Calcium	mg/L	12.81	22.22	0.2	244	<0.05
Magnesium	mg/L	3.27	6.6	0	109	<0.05
Sodium	mg/L	110.31	111.1	1.3	748	<0.05
Potassium	mg/L	3.56	4.91	0.1	34	<0.05
TDS	mg/L	2.42	0.3	1.54	3.27	<0.05
Total Alkalinity	mg/L	216.99	193.25	0	860	<0.05
Total Hardness	mg/L	1.32	0.55	-0.24	3.02	0.14
Carbonate	mg/L	5.75	19.57	0	372	<0.05
Temperature	°C	24.46	4.16	16	37	<0.05
Fluoride	mg/L	0.26	0.27	0	1.6	<0.05
Nitrate	mg/L	5.78	20.34	0	270	<0.05
Well Depth	ft	722.71	611.4	12	2,355.00	<0.05
Chloride	mg/L	33.19	62.86	2	665	<0.05
Sulfate	mg/L	22.72	40.71	0	500	<0.05
pH	-	7.65	1.11	3.2	10.8	<0.05

Results and Discussion

The 15 parameters analysed were: Calcium, Magnesium, Sodium, Potassium, TDS (log₁₀), Total Alkalinity, Total Hardness (log₁₀), Carbonate, Temperature, Fluoride, Nitrate, Well Depth, Chloride, Sulphate, and pH. These represent the major ions governing mineralisation, hardness, salinity, and the corrosive character of water, together with the principal species of direct toxicological and dental concern (fluoride, nitrate, pH). Government guideline values provide the benchmarks against which health relevance is judged: the WHO guideline for fluoride is 1.5 mg/L and for nitrate 50 mg/L (as NO₃⁻); the U.S. EPA primary maximum contaminant levels (MCLs) are 4.0 mg/L for fluoride and 10 mg/L for nitrate (as nitrogen), while secondary aesthetic standards are 250 mg/L for chloride, 250 mg/L for sulphate, and 500 mg/L for TDS, with an optimal pH range of 6.5–8.5 [14,16-18]. It should be noted that nitrate in this study is reported as the nitrate ion (NO₃⁻); the WHO guideline of 50 mg/L is expressed as nitrate ion, whereas the U.S. EPA MCL of 10 mg/L is expressed as nitrate-nitrogen (equivalent to approximately 45 mg/L as nitrate ion). Sodium has no federal primary MCL; it is managed aesthetically, with advisory guidance around 200 mg/L, above which taste is affected and sodium-restricted individuals may be at risk [2].

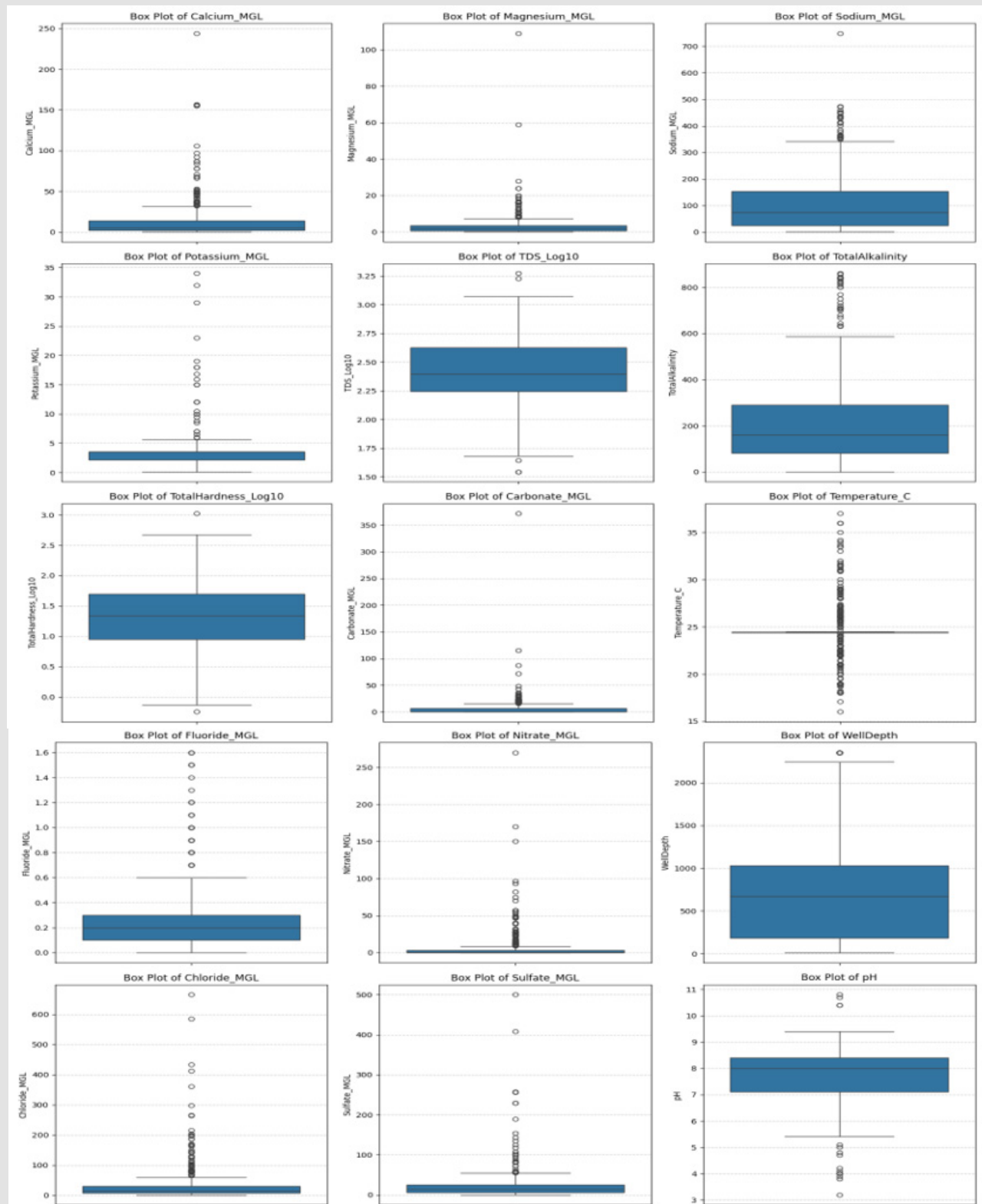


Figure 1: Boxplot (Whisker) for 15 Parameters.

Figure 1 presents box-and-whisker plots for the 15 parameters, displaying the median, interquartile range, whiskers, and outlier observations for each variable. Because the parameters span disparate units and scales, the plots reveal several shared features: pronounced right-skewness for Sodium, Nitrate, Chloride, Sulphate, and Calcium, each showing a compact central box with an extended upper whisker and numerous high-value outliers; a comparatively narrow, symmetric distribution for pH and Temperature; and log-transformed distributions for TDS and Total Hardness that compress the long upper tail. The outlier-rich parameters are precisely those whose maxima far exceed drinking-water guidelines, indicating that sub-populations of wells experience contamination events of health concern even when the median is low. Among the 15 parameters, the four with the greatest implications for oral, dental, and systemic health are

1. Nitrate
2. Fluoride
3. Sodium
4. Chloride

Nitrate (mean 5.78 mg/L; max 270 mg/L) shows the most extreme right-skewed boxplot relative to the EPA 10 mg/L (as nitrate-nitrogen) standard; its outlier mass reflects agricultural contamination, the leading driver of exceedance in regional aquifers, and denotes methaemoglobinaemia and reproductive risk [17,19]. Fluoride (max 1.6 mg/L) straddles the WHO 1.5 mg/L guideline; while optimal exposure prevents caries, values at or above the guideline carry dental-fluorosis risk, placing fluoride at the direct interface of dental benefit and harm [20,21]. Sodium (mean 110.31 mg/L; max 748 mg/L) exhibits the largest absolute dispersion; chronic high-sodium water contributes to hypertension and cardiovascular disease burden [2]. Chloride (max 665 mg/L) exceeds the 250 mg/L aesthetic standard in a subset of wells and accelerates corrosion of plumbing, potentiating metal release [14].

Compared with previous studies, the present findings are consistent with regional assessments showing that nitrate exceedances

are concentrated in shallow, unconfined aquifers beneath agricultural land, where a notable proportion of samples surpass the 10 mg/L (as nitrate-nitrogen) standard [19]. They also accord with global reviews demonstrating that community water fluoridation reduces caries prevalence but that dose-response is non-linear and fluorosis risk rises above 1.5 mg/L [20,21]. The novel contribution of the present boxplot analysis is the simultaneous, scale-comparable visualisation of all 15 parameters within a single frame, which exposes the shared right-skewed, outlier-driven structure of the health-relevant ions—a feature masked when parameters are examined in isolation. Figure 2 displays the frequency-distribution histograms for the 15 parameters. The distributions are characterised by unimodal, right-skewed shapes for most ionic parameters, with the bulk of observations concentrated at low values and a long right tail of high-concentration wells. TDS and Total Hardness, log-transformed, approximate more symmetric, near-normal distributions; pH clusters tightly around near-neutral to mildly alkaline values (mean 7.65). Bimodality is not pronounced, indicating that the dataset predominantly reflects a single hydrogeochemical population with a contaminant-enriched tail rather than two geographically distinct source populations. The four parameters with the greatest oral-, dental-, and systemic-health implications identified from the frequency distributions are

1. Nitrate
2. Fluoride
3. Sodium
4. TDS

The nitrate histogram is sharply peaked near zero with an extended tail crossing the 10 mg/L (as nitrate-nitrogen) threshold, confirming that a meaningful minority of wells deliver water unsuitable for infants; this frequency structure mirrors the regional finding that a measurable share of minor-aquifer samples exceed the nitrate MCL [19].

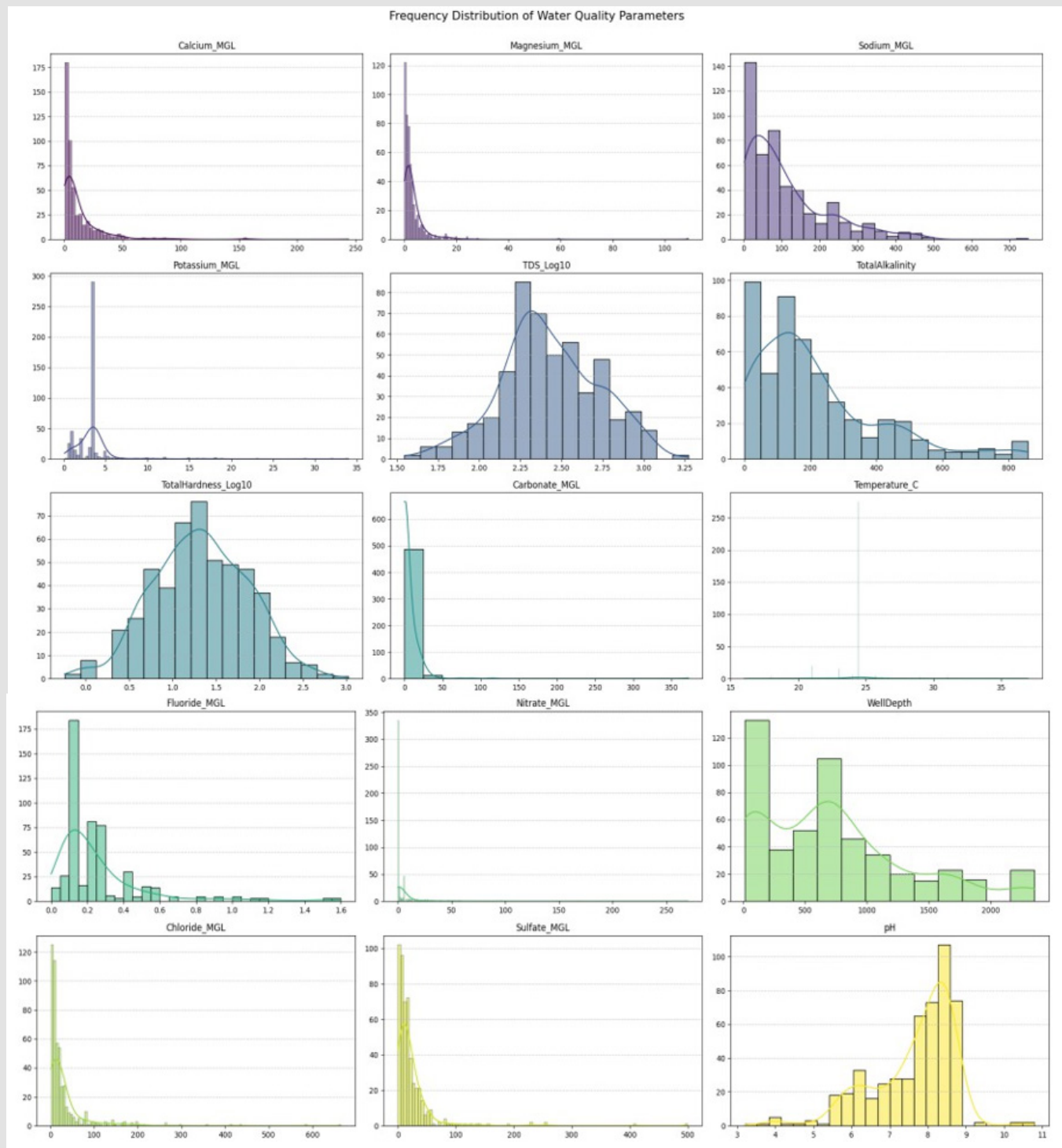


Figure 2: Frequency distributions for 15 parameters.

The fluoride distribution, though centred well below 1.5 mg/L, places a subset of wells within the caries- protective-to-fluorotic transition zone, underscoring the narrow therapeutic window [17,21]. Sodium and TDS show the broadest distributions, reflecting mineralisation and salinity gradients; elevated TDS is associated with taste rejection, scaling, and corrosivity that indirectly affects health by altering water consumption patterns and metal leaching [16]. Relative to earlier work, these distributions concur with WHO guidance that TDS and most major ions exert no direct toxic effect at typical concentrations but become health-relevant through aesthetic rejection and corrosion pathways at the high tail [16]. The distinguishing feature of the present analysis is the identification, within a unified distributional view, of a consistent contaminant-tail structure shared by the four health-relevant parameters—a pattern that justifies prioritising these ions for targeted monitoring and supports the geochemical

principle that salinity-related ions (Na, Cl, TDS) co-vary with land-use-driven contaminants such as nitrate. Figure 3 presents a Pearson correlation heatmap of the 15 parameters, with colour intensity encoding the strength and direction of pairwise relationships. The matrix is dominated by a block of strong positive correlations among the major ions—Sodium, Chloride, Sulphate, TDS, and Total Alkalinity—and a separate cluster linking Calcium, Magnesium, and Total Hardness. Weak or near-zero correlations characterise pairs such as pH–Well Depth and Fluoride–Nitrate, indicating largely independent geochemical controls. Negative correlations are sparse and modest, consistent with the absence of strong acid–base antagonism in the dataset. The two relationships carrying the greatest implications for oral, dental, and systemic health are

- 1. Calcium–Total Hardness (with Magnesium)
- 2. TDS–Sodium (with Chloride)

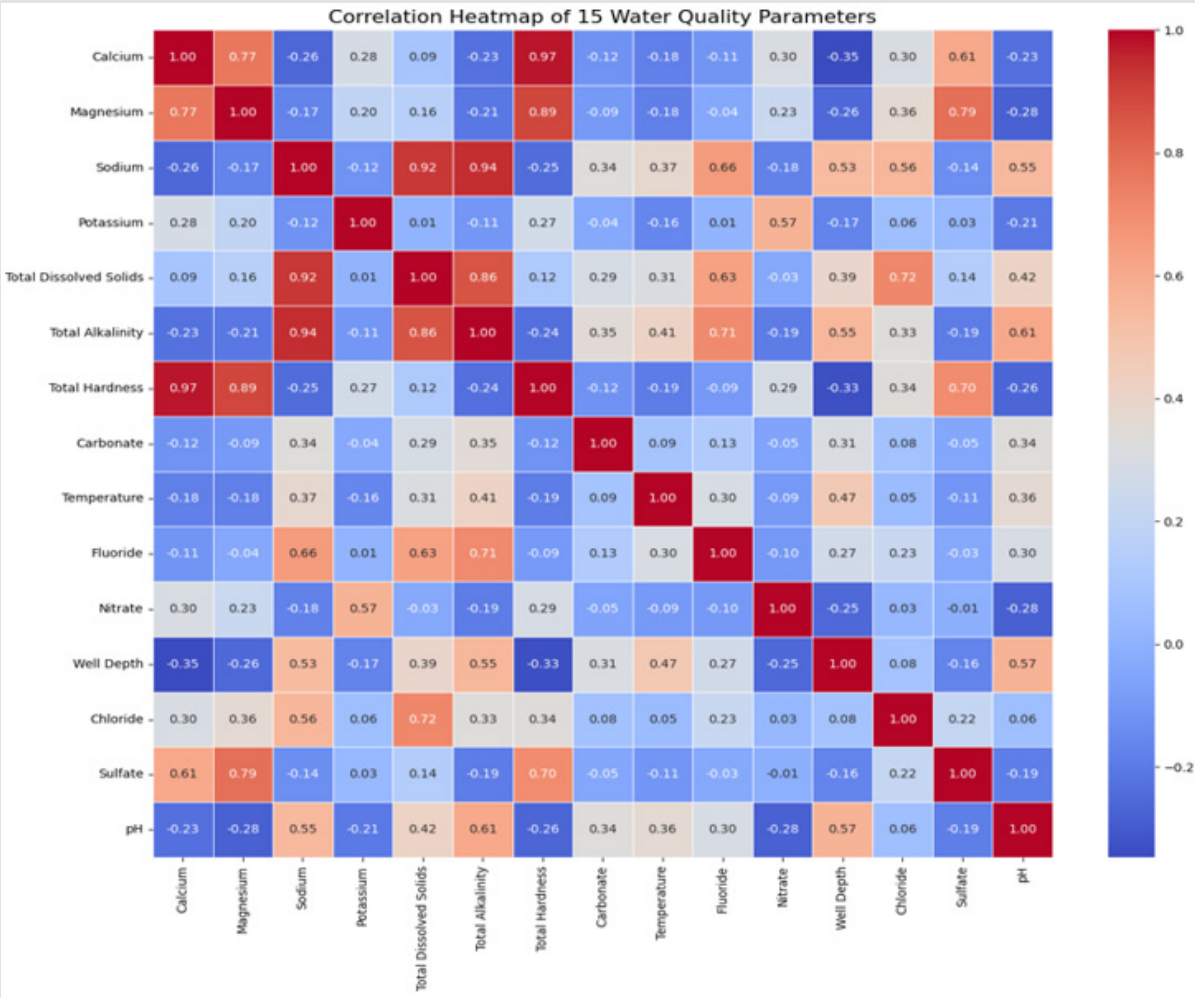


Figure 3: Correlation Heatmap.

The strong positive Calcium–Hardness–Magnesium cluster reflects the dominant contribution of these cations to water hardness; hardness is mechanistically relevant to dental and cardiovascular health because calcium and magnesium participate in enamel remineralisation and have been associated with protective cardiovascular effects, while very high hardness degrades palatability and promotes scaling [16]. The strong positive TDS–Sodium–Chloride relationship reflects salinity co-occurrence; because these ions share the same dissolution and agricultural sources, an elevated value of one flags potential elevation of the others, compounding hypertension and corrosion risk [2,14]. In comparison with previous studies, the observed correlation structure is consistent with the well-established geochemical grouping of salinity-related ions and hardness-related cations reported in regional groundwater assessments and ML groundwater-quality reviews [22-33]. The agreement in positive Na–Cl–TDS coupling is expected given their shared evaporite and agricultural provenance. The principal difference is that the present heatmap integrates all 15 parameters simultaneously, revealing that the health-relevant correlations are confined to two coherent geochemical blocks; this supports the scientific inference that targeted monitoring of a small set of indicator ions (hardness, sodium, chloride, TDS) can proxy for broader water-quality surveillance relevant to oral and systemic health.

Conclusion

This study applied a machine learning-assisted exploratory framework to a uniquely long-term, 86-year, 238-well, 505-sample groundwater chemistry dataset with the aim of identifying chemical factors relevant to oral and dental health. Using Google Colab-based statistical and visualization workflows, 15 physicochemical parameters were characterized through descriptive statistics, boxplot and frequency-distribution analysis, and correlation-matrix modeling. The results identified fluoride, nitrate, pH, and total dissolved solids as the parameters most directly relevant to oral health, each displaying right-skewed distributions with substantial guideline exceedances in a subset of wells, while correlation analysis revealed strong geochemical linkages between fluoride and alkalinity and sodium that may compound fluorosis risk in specific hydrogeological settings. These findings were broadly consistent with prior fluoride- and nitrate-focused groundwater research, while also revealing greater distributional heterogeneity and a distinct fluoride-nitrate source pattern attributable to the extended temporal and spatial coverage of the dataset. Overall, this study demonstrates that accessible, reproducible machine learning tools can meaningfully support the exploratory screening of environmental chemical exposures relevant to oral health, providing a foundation for future predictive modeling, targeted water-quality monitoring, and public health interventions aimed at reducing oral disease burden in populations reliant on groundwater sources.

References

1. Barranca-Enríquez A, Romo-González T (2022) Your health is in your mouth: A comprehensive view to promote general wellness. *Front Oral Health* 3: 971223.
2. (2022) World Health Organization. Global oral health status report: Towards universal health coverage for oral health by 2030. World Health Organization, Geneva, Switzerland.
3. Aldhyani THH, Yaari MA, Alkahtani H, Maashi M (2020) Water quality prediction using artificial intelligence algorithms. *Appl Bionics Biomech* 2020: 6659314.
4. Azrouz M, Mabrouki J, Fattah G, Guezzaz A, Aziz F (2022) Machine learning algorithms for efficient water quality prediction. *Model Earth Syst Environ* 8(2): 2793-2801.
5. Zhu JJ, Yang M, Ren ZJ (2023) Machine learning in environmental research: Common pitfalls and best practices. *Environ Sci Technol* 57(46): 17671-17689.
6. Podgorski J, Berg M (2022) Global analysis and prediction of fluoride in groundwater. *Nat Commun* 13(1): 4232.
7. Mohammed MAA, Mohamed A, Szabó NP, Szűcs P (2024) Development of machine-learning-based models for identifying the sources of nitrate and fluoride in groundwater and predicting their human health risks. *Int J Emerg Water Res* 8: 161-180.
8. Rani JPA, Nivasini R, Rubavathi CY, Jona P (2025) Fluoride risk prognosis: A pioneering ensemble machine learning approach for groundwater contamination prediction in parts of the East Coast of India. *Water* 17(6): 909.
9. Kang IA, Ngnamsie Njimboum S, Lee KO, Kim JD (2022) DCP: Prediction of dental caries using machine learning in personalized medicine. *Appl Sci* 12(6): 3043.
10. Qu X, Zhang C, Houser SH, Zhang J, Zou J, et al. (2022) Prediction model for early childhood caries risk based on behavioral determinants using a machine learning algorithm. *Comput Methods Programs Biomed* 227: 107221.
11. Toledo Reyes L, Knorst JK, Ortiz FR, Brondani B, Emmanuelli B, et al. (2023) Early childhood predictors for dental caries: A machine learning approach. *J Dent Res* 102(9): 999-1006.
12. Wang X, Zhang P, Lu H, Luo D, Yang D, et al. (2025) Risk prediction models for dental caries in children and adolescents: A systematic review and meta-analysis. *BMJ Open* 15(3): e088253.
13. (2024) Texas Water Development Board. Groundwater Database. Austin: Texas Water Development Board.
14. (2024) U.S. Environmental Protection Agency. 2024 Edition of the Drinking Water Standards and Health Advisories Tables. Washington, D.C.: U.S. EPA; Report No.: EPA 822-S-24-001.
15. (2022) World Health Organization. Chemical fact sheets: Hardness. In: *Guidelines for Drinking-water Quality*. (4th Edn.), Geneva: World Health Organization.
16. (2022) World Health Organization. Chemical fact sheets: Total dissolved solids. In: *Guidelines for Drinking-water Quality*. 4th ed. Geneva: World Health Organization.
17. (2022) World Health Organization. guidelines for drinking-water quality: Fourth edition incorporating the first and second addenda. Geneva: World Health Organization.

18. (2026) Texas Commission on Environmental Quality. Drinking Water Standards Governing Drinking Water Quality and Reporting Requirements for Public Water Systems. 30 TAC Chapter 290, Subchapter F. Austin: TCEQ.
19. Reedy RC, Scanlon BR (2023) Groundwater quality in Texas: Evaluation of major and minor aquifers. Austin: Bureau of Economic Geology, The University of Texas at Austin.
20. Ejiofor ZI, Walsh T, Lewis SR, Riley P, Boyers D, et al. (2024) Water fluoridation for the prevention of dental caries. *Cochrane Database Syst Rev* 10(10): CD010856.
21. Nascimento GG, Peres MA, Do LG (2025) Water fluoridation and oral health equity: A systematic review. *J Dent Res* 104(2): 123-132.
22. Egbueri JC, Omeka ME, Agbasi JC (2023) Geochemical evaluation and machine learning prediction of groundwater suitability for drinking and irrigation. *Environ Earth Sci* 82(14): 345.
23. Elsayed S, Hussein H, Elsherbiny O, Farouk M (2023) Predicting water quality index in arid regions using machine learning algorithms. *Water Resour Manage* 7(8): 3125-3142.
24. Xiong Y, Liu D, Zhou Y (2023) Hydrochemical characteristics and water quality assessment of groundwater using machine learning methods. *Chemosphere* 311: 137021.
25. Ahuja S (2021) *Handbook of Water Purity and Quality*. (2nd Edn.), Academic Press, London, United Kingdom.
26. Azrour M, Mabrouki J, Fattah G, Guezzaz A, Aziz F (2022) Machine learning algorithms for efficient water quality prediction. *Model Earth Syst Environ* 8(2): 2793-2801.
27. Boehmer TJ, Lesaja S, Espinoza L, Ladva CN (2023) Community water fluoridation levels to promote effectiveness and safety in oral health—United States, 2016–2021. *MMWR Morb Mortal Wkly Rep* 72(22): 593-596.
28. (2024) Health Canada. Guidelines for Canadian drinking water quality: Guideline technical document for public consultation—operational parameters: Calcium, magnesium, hardness, chloride, sulphate, total dissolved solids and hydrogen sulphide in drinking water. Health Canada, Ottawa, Canada.
29. Lee JM, Ko KS, Yoo K (2023) A machine learning-based approach to predict groundwater nitrate susceptibility using field measurements and hydrogeological variables in the Nonsan stream watershed, South Korea. *Appl Water Sci* 13(12): 242.
30. Picetti R, Deeney M, Pastorino S, Miller MR, Shah A, et al. (2022) Nitrate and nitrite contamination in drinking water and cancer risk: A systematic review with meta-analysis. *Environ Res* 210: 112988.
31. (2026) Texas Water Development Board. Groundwater database. Texas water development board, Austin, Texas, United States.
32. (2022) World Health Organization. Chemical fact sheets: Total dissolved solids. Guidelines for drinking- water quality. (4th Edn.), incorporating the 1st and 2nd addenda), World Health Organization, Geneva, Switzerland.
33. (2022) World Health Organization. Chemical fact sheets: Hardness. Guidelines for drinking-water quality. (4th Edn.), incorporating the 1st and 2nd addenda), World Health Organization, Geneva, Switzerland.

ISSN: 2574-1241

DOI: 10.26717/BJSTR.2026.67.010460

Kevin Heo . Biomed J Sci & Tech Res



This work is licensed under Creative Commons Attribution 4.0 License

Submission Link: <https://biomedres.us/submit-manuscript.php>



Assets of Publishing with us

- Global archiving of articles
- Immediate, unrestricted online access
- Rigorous Peer Review Process
- Authors Retain Copyrights
- Unique DOI for all articles

<https://biomedres.us/>