Bayesian Network-Based Water Quality and Disease Prediction in Sindh

Fatima Dossa
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan

Maria Adnan
Dhanani School of Science and Engineering
Habib University
Karachi, Pakistan

Syed Nisar Hussain
Dhanani School of Science and Engineering
Habib Univeristy
Karachi, Pakistan

***Abstract*—Waterborne diseases remain a significant public health challenge in Sindh, Pakistan, where communities are heavily relying on hand-pump water sources that are often contaminated with chemical, microbial, and physical pollutants. These diseases, including cholera, malaria, and skin infections, pose severe health risks, particularly in rural areas where access to clean water and effective monitoring systems is limited. Our project addresses this critical issue by developing a predictive model using Bayesian Networks to assess the likelihood of disease occurrences based on water quality data from Sindh. The data set, which includes contamination metrics from 90 water samples collected in three districts – Badin, Sanghar, and Sukkur– under the study named *Assessment of Water Quality of Hand Pumps from Lower, Middle and Upper Districts of Sindh, Pakistan* forms the basis of the analysis. Nodes in the Bayesian Network represent villages, water quality parameters, and diseases, interconnected to model causal relationships. Probabilities are derived from primary data collected during the mentioned study. By integrating probabilistic graphical models with localized data to produce a predictive model, our aim is to provide information on disease prevention and inform policy interventions to improve public health in vulnerable communities.**

***Index Terms*—Bayesian Networks, Diseases, Water, Prediction**

I. Introduction

Access to safe and clean water is a fundamental human need, yet many regions, including Pakistan, struggle with contamination, posing risks like waterborne diseases for vulnerable communities. Waterborne diseases are a persistent public health issue, particularly in developing regions such as Sindh, Pakistan, where communities often depend on untreated hand pump water for drinking and domestic use. Contaminated water is one of the primary causes of the transmission of numerous diseases, including cholera, typhoid, hepatitis, and parasitic infections. The health impact of these diseases is significant, with vulnerable populations, such as children and the elderly, being disproportionately affected.

Despite their severity, systematic and data-driven approaches to identifying and addressing the risks of waterborne diseases in these areas remain limited. **Predicting water quality and disease risk** is vital to public health. Traditional methods have often focused on contamination mapping and microbial assessments; however, newer approaches such as **machine learning and probabilistic models** are proving essential not only for assessing contamination, but also for predicting disease outcomes.

This project targets the need for predictive tools that can assess the risk of waterborne diseases in regions like Sindh. The goal is to develop a Bayesian Network model capable of mapping the causal relationships between water quality parameters and disease occurrences. Bayesian Networks are probabilistic graphical models that provide a robust framework for reasoning under uncertainty, making them particularly suitable for applications in public health. By representing variables as nodes and their dependencies as directed edges, Bayesian Networks can help capture the intricate causal relationships between various water quality metrics and their potential health impacts.

The dataset utilized in this study is derived from the research conducted during the Summer Tehqiq Research Program (STRP) at Habib University. It comprises water quality data collected from 90 samples across three districts in Sindh: Badin (lower district), Sanghar (middle district), and Sukkur (upper district). One of the most significant contributions of this project is its localized approach to disease risk prediction. While there is extensive research on waterborne diseases and water quality, very few studies have focused on the specific context of Sindh, Pakistan. This project fills that gap by integrating locally sourced data into a probabilistic framework, enabling precise and context-aware predictions. Furthermore, by leveraging Bayesian Networks, this model allows for the visualization and analysis of causal relationships, which can be invaluable for policymakers and health practitioners.

II. Literature Review

Recent studies have explored diverse methodologies, from QMRA to machine learning algorithms and Bayesian geostatistical techniques, each providing analytical data on water quality and health risks. The following literature review examines how these methods have been incorporated into previous works, and identifies whether there is any research gap, hence offering the **value of Bayesian network models (and our project)** in this field.

Research published in the *International Journal of Environmental and Public Health* did an assessment of water quality and disease prediction—similar to our project—in different districts of Sindh, Pakistan. Mainly focusing on primary school children, the researchers used a multistage **random sampling technique** across 425 primary schools, gathering data on water contamination sources and levels to build risk profiles by region within Sindh **[1]**. They employed a rigorous **Quantitative Microbial Risk Assessment (QMRA) model** to estimate the risk of infection among those primary schoolchildren. For **water quality analysis**, this research adhered to **WHO microbial water quality standards**, focusing on a variety of pathogens (E. coli, Salmonella spp., Shigella, V. cholerae, etc.) in the drinking water of selected schools. The QMRA model incorporated parameters such as contamination levels and infection probabilities, applying statistical methods to calculate the annual probability of infection and subsequent illness for various pathogens.

For **disease prediction**, the study relied on **hardcore mathematical computation**; the **Risk Characterization Equation**, a standardized mathematical framework within the QMRA approach which provides quantitative estimates of infection risks based on contamination levels, was used to compute both the probability of infection and probability of illness by combining microbial test data with infection parameters. This model, grounded in probabilistic assessments, has shown effectiveness in estimating infection risks across varied contaminants and populations. Some of the key findings of this research which are relevant and may prove beneficial for our project are:

1) High Prevalence of Contaminants: Approximately half of the water samples across Sindh's primary schools were found to be contaminated with E. coli, Salmonella spp., Vibrio cholerae, and Shigella.
2) Geographical Variation in Risk: The risk of infection varied significantly across regions. South Sindh, heavily reliant on surface water sources, exhibited the highest contamination levels and infection risks. Notably, schools in Karachi had the highest probability of illness from Campylobacter and Rotavirus.
3) Correlations Between Pathogens: The study found that Vibrio cholerae contamination correlated with Salmonella spp., Campylobacter, and Rotavirus, suggesting overlapping sources of contamination or similar exposure pathways for these pathogens.
4) Annual Infection Risk Estimates: The probability of schoolchildren contracting infections due to bacterial contamination was notably high, with over 94% of children at risk from bacteria like E. coli, Shigella, and Salmonella spp. The highest illness risks were associated with Campylobacter, followed by Salmonella spp., E. coli, and Shigella.

Another study used a **Machine Learning Algorithm** for **timely detection of fish diseases by analyzing water quality**; the researchers used algorithms like **Gradient Boosting**, which utilize a decision-tree-based structure optimized through regression. After training, the machine learning model was able to predict future Water Quality Index (WQI) values and identify potential diseases associated with changes in water quality parameters. **The relationship between the year of data collection and the WQI value indicated that over the years, the WQI values increased, suggesting an improvement in water quality over time [2]**; the "Predicted WQI" showed the forecast for the same water sources four months into the future. The model was trained using data collected from 500 sources, with three samples taken at four-month intervals from each source. On average, their model could predict disease with an impressive 92% accuracy. However, this study focused more on **regression-based optimization rather than probabilistic graphical modeling**.

We also found a study in which the researchers had an objective almost similar to ours: use of **Bayesian Networks in predicting contamination of drinking water** with E. coli in rural Vietnam. Even though this research was not set in Pakistan, this is the only study we could find which has used Bayesian Networks as a prediction model. The model incorporated a wide range of variables to assess their collective impact on water quality. One of the key techniques used in the study is the **membrane filtration method** to measure the presence and concentration of E. coli in water samples **[3]**. This microbial analysis provided Colony-Forming Unit (CFU) count, which served as a direct measure of contamination levels. These CFU counts were then integrated into the BN model as one of the key indicators of water quality. This study specifically made use of a Bayesian Network model which was developed using **Netica** software, a tool designed for constructing and analyzing Bayesian networks. The model was refined using **Bayesian inference**, which updates the probability estimates for different outcomes as more data is introduced. This iterative process ensured that the predictions remain accurate as new information becomes available.

A similar research used two **geostatistical techniques, Bayesian kriging and classical Ordinary Kriging, to predict groundwater contamination** in Jhelum, Pakistan **[4]**. Both techniques were used to assess contamination levels by focusing on six critical water quality parameters, which are essential indicators of water quality, and deviations from standard levels signal potential contamination that poses risks to public health. However, this research focused mainly on **spatial analysis**, rather than direct disease prediction.

Even though this study did not use Bayesian networks itself, it did apply Bayesian kriging, which is an advanced form of geostatistical modeling that incorporates Bayesian inference to improve prediction accuracy. While not a full graphical model like a Bayesian network, Bayesian kriging used the probabilistic nature of Bayesian statistics to incorporate prior information, making it more effective in handling uncertainties in spatial contamination data compared to classical kriging. The study thus demonstrated how Bayesian kriging, supported by MCMC simulations, provides better predictions, especially in areas with limited data, such as the outskirts of Jhelum.

One **research gap** which favors the importance of our project is that this study stops short of disease prediction, **primarily focusing on contaminant mapping and hotspot identification**. Bayesian kriging's probabilistic capabilities are strong for contamination predictions, but they are not designed to establish **multivariate connections between water quality metrics and potential health outcomes**. Despite leveraging Bayesian inference, these methods are **mathematically oriented** and do not integrate graphical models that map relationships across multiple variables—an essential component of our project. Moreover, while all the mentioned studies have numerous impressive models for water quality analysis exist, few mention disease prediction, and those that do often rely on generalized WHO data. **Only one study has used a machine learning algorithm for prediction in Pakistan, and none have applied Bayesian Networks for disease prediction, particularly in Pakistan**. Hence, this project represents a major **breakthrough**, in which we build upon the insights from all the above cited papers by utilizing **firsthand data collected by our own university's researchers** under the STRP; this data from the research report serves as the empirical foundation of our **Bayesian Network**, enabling us to **model the causation and relationships between water quality parameters and disease likelihood in a way that previous methodologies have not fully captured.**

The STRP paper analyzes the water quality of hand pump water from 3 different districts of Sindh; the water in these areas is contaminated with harmful microbes and chemicals, with floods heightening this issue by damaging the existing water pumps. The water samples are collected from the **lower, middle, and upper districts of Sindh—namely, Badin, Sukkur, and Sanghar**, respectively. In each district, a total of 30 samples were collected from six different villages, with five samples taken from separate households within each village **[5]**.

These samples were then tested chemically, microbially, and some underwent arsenic testing. In the chemical testing, **16 water quality parameters** were analyzed. For microbial testing, the presence of E. coli and Coliform bacteria was assessed and quantified using the **Most Probable Number (MPN) method**. The water quality in these districts is severely compromised by a mix of environmental, infrastructural, and human factors. The prevalence of coliform in Badin, nitrate contamination in Sukkur, and high TDS levels in Sanghar along with other elevated water quality parameters present serious public health risks which lead to a multitude of diseases identified in the paper. **We will extend this research by developing a Bayesian Network Model, assigning prior and conditional probabilities based on the extracted data to enhance disease prediction.**

III. Methodology

A. *Construction of the Model:*

The methodology that we employed to reach the goal of disease prediction through a Bayesian network by assessing the water quality in Sindh can be broken down into several key steps.

The preliminary research assessed water quality through a total of 30 samples from each district; 5 samples from each of the 6 villages in that district. Each sample was chemically and microbially tested and evaluated based on 19 water quality parameters: *Total Alkalinity (TA), pH, Cyanuric Acid (CYA),* (CO_3^{2-})*, Hardness* $(CaCO_3)$*, Total Chlorine* (Cl_2)*, Free Chlorine* (FCl_2)*, Free Bromine* (Br_2)*, Nitrate* (NO_3^-)*, Nitrite* (NO_2^-)*, Iron (Fe), Chromium* (Cr^{6+})*, Lead (Pb), Copper (Cu), Mercury (Hg), Fluoride* (F^-)*, E.Coli and Coliforms* and this data was recorded and maintained in an Excel sheet.

Since the scope of this project was **disease inference based on the different water quality parameters**, we eliminated those parameters from our model which only led to short-term effects rather than diseases like eye burn or skin rash. The diseases caused by high levels of the selected parameters were found through credible platforms like WHO, PubMed, and CDC.

The model has the initial root node named **"Districts of Sindh"** with 3 states: Sukkur, Sanghar and Badin which we are terming as Level 0. This node has edges to all the nodes of the shortlisted water quality parameters (Level 1) which in turn cause different diseases, also represented as separate nodes (Level 2). Each node has its own CPT and the model is as follows:

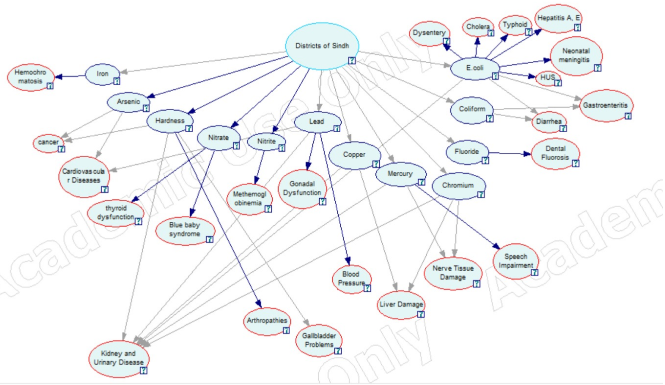


Fig. 1. BN model for Water Quality and Disease Prediction

B. *Assigning Probabilities:*

The probability for each of the three states in the initial node was assigned equal likelihood which is 1/3.

To assign the conditional probability for each of the parameters in level 1,

$$P(parameter = high|district = Sukkur)$$

$$P(parameter = high|district = Sanghar)$$

$$P(parameter = high|district = Badin)$$

we made use of the data of the preliminary research data. We divided their main sheet into three separate sheets for each district. For each sample testing result, according to the values obtained for each parameter, we assigned 1(high) or 0(optimum) on Excel according to where the value falls in the WHO and testing kit specified threshold for that parameter. The probabilities for each district were calculated based on the count of 1's and the total number of samples (30). To address instances of zero probability, Laplacian smoothing was applied. The formula was as follows:

$$P(\text{parameter} = \text{high} \mid \text{district} = \text{Sukkur})$$

$$= \frac{\text{Count of 1's of parameter in Sukkur} + 1}{30 + 2}$$

Alternatively, we also generated the same probabilities through a python program to verify the ones generated through Excel and they were both accurate. For the level 2 probabilities, where we had to determine

$$P(disease = yes|parameter = high)$$

and

$$P(disease = yes|parameter = optimum)$$

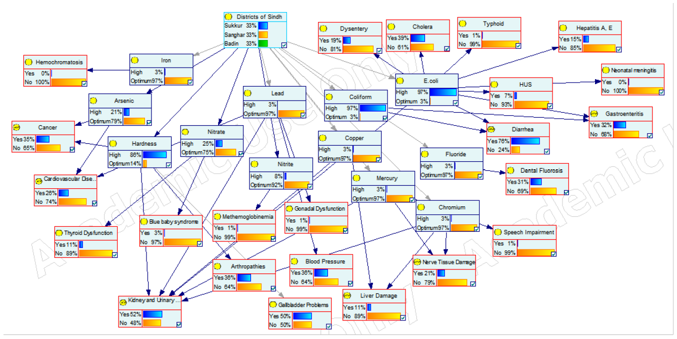
there was no data set available which had the parameters and the diseases both, so we had to take a qualitative approach and consult the domain experts.

IV. Results/Findings

A. General Observations:

Once we had made the model and had identified prior and conditional probabilities of all nodes from their respective sources, we began to observe changes, patterns and prevalent aspects of causality in the BN. Without any evidence on the model, we noticed the following observations:

1) Some of the impurities which are more commonly found in water of the 3 districts include *E.Coli (97%), Coliform (97%), Hardness (86%), Nitrate (25%), Arsenic (21%) and Nitrite (8%).*
2) Because of the parameters mentioned above, the diseases caused by them are consequently more prevalent in the areas too. These diseases include *Dysentery (19%), Cholera (39%), Cancer (35%), Cardiovascular Diseases (26%), Kidney and Urinary Diseases (50%), Arthopathies (36%), Blood pressure (36%), Gallbladder problems (50%), Dental Fluorosis (31%), Diarrhea (76%), and Gastroenteritis (32%).*



Districts of Sindh Sukkur 33% Sanghar 33% Badin 33%
Dysentery Yes 19% No 81%
Cholera Yes 39% No 61%
Typhoid Yes 1% No 99%
Hepatitis A, E Yes 15% No 85%
Hemochromatosis Yes 0% No 100%
Iron High 3% Optimum97%
E.coli High 97% Optimum 3%
HUS Yes 7% No 93%
Neonatal meningitis Yes 0% No 100%
Arsenic High 21% Optimum79%
Lead High 3% Optimum97%
Coliform High 97% Optimum 3%
Gastroenteritis Yes 32% No 68%
Cancer Yes 35% No 65%
Hardness High 86% Optimum14%
Nitrate High 25% Optimum75%
Copper High 3% Optimum97%
Diarrhea Yes 76% No 24%
Fluoride High 3% Optimum97%
Nitrite High 8% Optimum92%
Dental Fluorosis Yes 31% No 69%
Cardiovascular Dse.. Yes 26% No 74%
Mercury High 3% Optimum97%
Methemoglobinemia Yes 1% No 99%
Gonadal Dysfunction Yes 1% No 99%
Chromium High 3% Optimum97%
Blue baby syndrome Yes 3% No 97%
Thyroid Dysfunction Yes 11% No 89%
Speech Impairment Yes 1% No 99%
Arthropathies Yes 36% No 64%
Blood Pressure Yes 36% No 64%
Nerve Tissue Damage Yes 21% No 79%
Kidney and Urinary... Yes 52% No 48%
Gallbladder Problems Yes 50% No 50%
Liver Damage Yes 11% No 89%

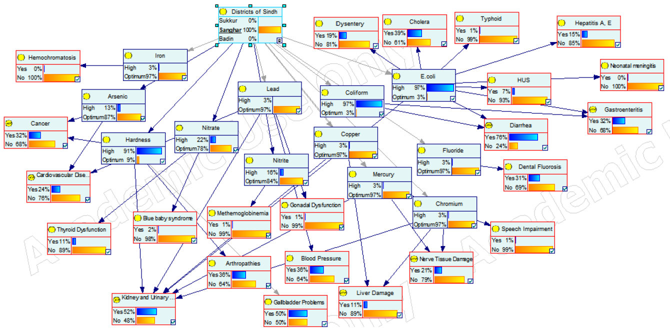
Fig. 2. Major Water Quality Parameters and Diseases Distribution in All Districts Without Evidence

This data depicted in the figure seemed to be in sync with the data from the STRP research.

B. Data Verification by Evidence Instantiation on Level 0:

We continued forward by instantiating evidences on the root node *Districts of Sindh* by putting Sukkur, Sanghar, and Badin = 1 one-by-one to note changes in the water quality parameters. This was done to confirm and validate the individual district-parameter relationship from the STRP data.

1) Sanghar – upon providing evidence – showed an increase of 5% in Hardness, 3% in Nitrate levels and 8% in Nitrite levels, with E.Coli and Coliform still being very high.



Districts of Sindh Sukkur 0% Sanghar 100% Badin 0%
Dysentery Yes 19% No 81%
Cholera Yes 39% No 61%
Typhoid Yes 1% No 99%
Hepatitis A, E Yes 15% No 85%
Hemochromatosis Yes 0% No 100%
Iron High 3% Optimum97%
E.coli High 97% Optimum 3%
HUS Yes 7% No 93%
Neonatal meningitis Yes 0% No 100%
Arsenic High 13% Optimum87%
Lead High 3% Optimum97%
Coliform High 97% Optimum 3%
Gastroenteritis Yes 32% No 68%
Cancer Yes 32% No 68%
Hardness High 91% Optimum 9%
Nitrate High 22% Optimum78%
Copper High 3% Optimum97%
Diarrhea Yes 76% No 24%
Fluoride High 3% Optimum97%
Nitrite High 16% Optimum84%
Dental Fluorosis Yes 31% No 69%
Cardiovascular Dse.. Yes 24% No 76%
Mercury High 3% Optimum97%
Methemoglobinemia Yes 1% No 99%
Gonadal Dysfunction Yes 1% No 99%
Chromium High 3% Optimum97%
Blue baby syndrome Yes 2% No 98%
Thyroid Dysfunction Yes 11% No 89%
Speech Impairment Yes 1% No 99%
Arthropathies Yes 36% No 64%
Blood Pressure Yes 36% No 64%
Nerve Tissue Damage Yes 21% No 79%
Kidney and Urinary... Yes 52% No 48%
Gallbladder Problems Yes 50% No 50%
Liver Damage Yes 11% No 89%

Fig. 3. Evidence on Level 0: Districts of Sindh = Sanghar

2) Likewise, Sukkur showed a relatively high increase of percentages in Hardness, Nitrite, Nitrate, and Arsenic. E.Coli and Coliform still remain very high.

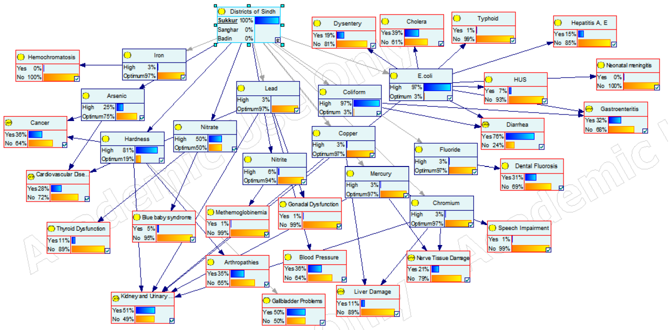


Fig. 4. Evidence on Level 0: Districts of Sindh = Sukkur

3) Lastly, Badin showed an increase of 2% in Hardness, and 4% in Arsenic. Again, as expected, E.Coli and Coliform still remain very high. However, one interesting observation was that Badin's Nitrate and Nitrite levels decreased significantly, with 5% decrease in Nitrite, 22% decrease in Nitrate levels.

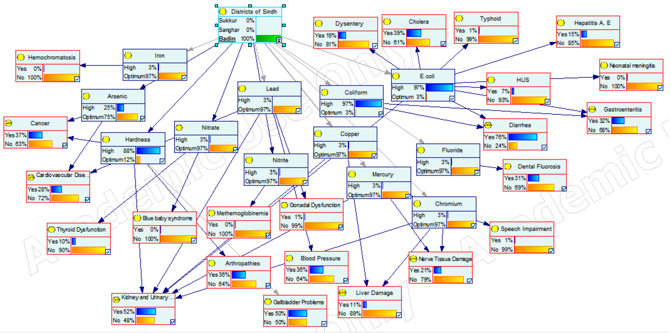


Fig. 5. Evidence on Level 0: Districts of Sindh = Badin

After cross-checking the changes in Conditional Probabilities from the MS Excel sheet provided by the STRP researchers, we then moved on to predict diseases given certain water quality parameters specific to district-wise context.

C. Disease Prediction by Contextual District-wise Evidence Instantiation on Level 1:

After providing evidence on the root node, we noted down the water quality parameters which were high (beyond the threshold) in that particular district, and then continued to provide evidence on only the above-identified nodes one-by-one. For example, when we instantiated Sanghar and observed Hardness, Nitrate, Nitrite, E.Coli and Coliform in high amounts in this district, we provided evidences on all 5 of the parameters individually to observe and predict potential diseases caused by them and their probabilities.

1) For Sanghar, providing evidence on Hardness predicted an increase in chances of Cancer by 1%. Similarly, Nitrate predicted an 8% increase in Blue Baby Syndrome and 2% increase in Thyroid Dysfunction; Nitrite predicted an increase of 8% in Methemoglobinemia; E.Coli and Coliform predicted 2-3% of increase in Dysentery, Cholera, Typhoid, HUS, Diarrhea, and Gastroenteritis.

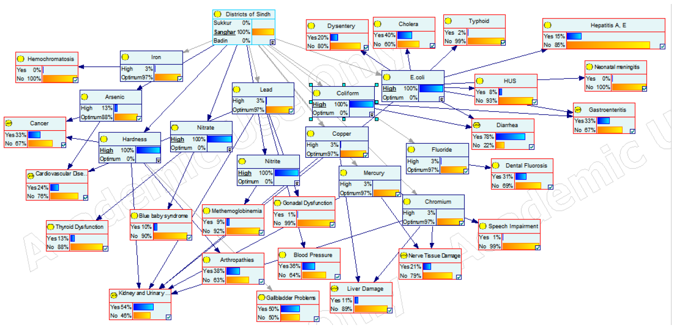


Fig. 6. Evidence on Level 0: Districts of Sindh = Badin

2) For Sukkur, providing evidence on Arsenic predicted an increase in chances of Cancer by 34% and Cardiovascular Diseases by 24%. Similarly, evidence on E.Coli and Coliform predicted a 2-3% increase in Dysentery, Cholera, Typhoid, HUS, Diarrhea, and Gastroenteritis; evidence on Hardness predicted a 3% increase in Cancer, Arthopathies, and Kidney Diseases; evidence on Nitrate predicted a 5% increase in Blue Baby Syndrome; and evidence on Nitrite predicted an 8% increase in Methemoglobinemia.

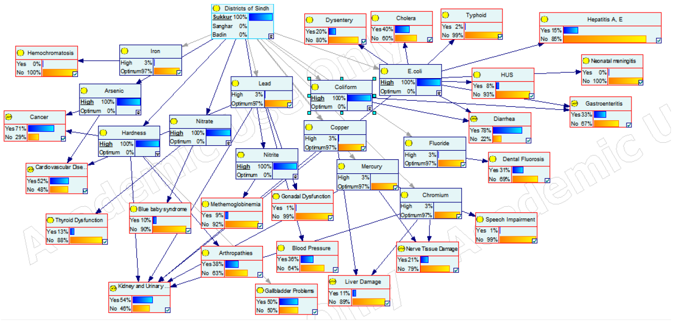


Fig. 7. Evidence on Level 0: Districts of Sindh = Badin

3) For Badin, providing evidence on Arsenic predicted an increase in chances of Cancer by 34% and Cardiovascular Diseases by 24%. Similarly, evidence on Hardness predicted a 1-2% increase in Cancer, Arthopathies, and Kidney Diseases; evidence on E.Coli and Coliform predicted a 2-3% increase in Dysentery, Cholera, Typhoid, HUS, Diarrhea, and Gastroenteritis.

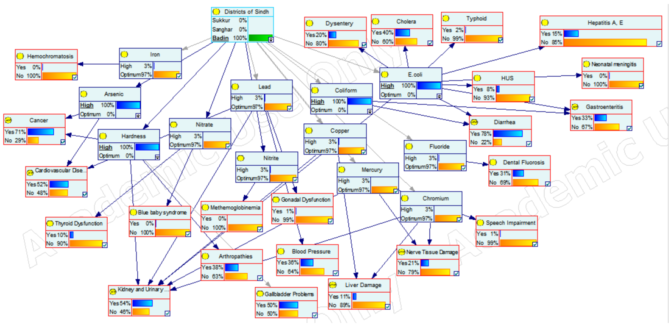


Fig. 8. Evidence on Level 0: Districts of Sindh = Badin

V. Discussion

Most diseases that show a high likelihood of occurrence after evidence is set like Kidney and Urinary diseases, Gallbladder Problems, Arthropathies, Cholera, Diarrhea, and Dysentery since the water quality parameters that cause them are "high" in majority or all the samples and common along the districts. For example, diseases like Diarrhea and Cholera are caused by bacteria like E coli or coliforms and all our samples from the earlier research for all three districts tested positive or "high" for both parameters since their MPN values were 1 or more and in safe water it should be less than 1. Thus, indicating that the diseases caused as a result are more likely to occur in these districts. Similarly, hardness also tests as "high" for majority of the samples across the three 3 districts, increasing the likelihood of diseases such as arthropathies and kidney and urinary disease.

The findings of the model revealed significant insights into the relationships between water quality parameters and disease prevalence across the three districts of Sindh. One of the most striking observations is the consistently high levels of *E.Coli* and *Coliform* across all districts, which directly cause a high likelihood of diseases such as *Diarrhea*, *Cholera*, *Typhoid*, *HUS*, and *Gastroenteritis*. These diseases are waterborne and arise from bacterial contamination, making them highly prevalent in areas with inadequate water treatment or poor sanitation. For instance, in Badin, Sukkur, and Sanghar, the majority of the water samples indicated MPN values greater than 1, a threshold beyond which water is considered unsafe. This strong causality emphasizes the urgent need for intervention to reduce microbial contamination in these districts.

In addition to bacterial contamination, other water quality parameters, such as *Arsenic* and *Hardness*, also demonstrated a profound impact on disease prediction. For instance, in *Sukkur* and *Badin*, evidence on Arsenic levels revealed a significant increase in the likelihood of *Cancer* and *Cardiovascular Diseases*, with probabilities rising by 34% and 24%, respectively. This aligns with global research on the carcinogenic and cardiovascular risks associated with prolonged exposure to high levels of Arsenic in drinking water. Similarly, Hardness showed a smaller yet notable increase in the probabilities of *Kidney Diseases*, *Arthropathies*, and *Cancer*, particularly in Sukkur and Sanghar. This highlights the long-term health risks posed by excessive mineral content in water, which can exacerbate chronic conditions over time.

Interestingly, the district-specific patterns in water quality parameters also influenced the likelihood of rare diseases. For example, high *Nitrate* levels in Sanghar and Sukkur predicted an increased risk of *Blue Baby Syndrome* and *Thyroid Dysfunction*, while elevated *Nitrite* levels predicted a higher probability of *Methemoglobinemia*. This underscores the value of localized water quality assessment for precise disease prediction. Notably, Badin exhibited a unique trend where *Nitrate* and *Nitrite* levels significantly decreased, contrasting with the patterns observed in the other two districts. This reduction contributed to a decreased likelihood of some health risks, highlighting potential variations in water sources or treatment methods within Badin.

These findings validate the use of probabilistic models for understanding disease causality in environmental contexts. The model not only aligns with prior data from STRP but also provides actionable insights into the prioritization of water quality interventions. Targeted measures, such as addressing microbial contamination and reducing Arsenic levels, could lead to substantial public health benefits in these regions.

VI. Limitations

Our BN model is subject to the following limitations.

- We have assumed the nodes with multiple parents to be Noisy OR structures, which simplifies the relationships of those nodes. Hence, this may not be able to capture all real-world complexities of dependencies on other factors.
- While we had domain experts verify the probabilistic outcomes of the model, expert representation was limited to one district-specific domain expert, and another general advisor. This may have potentially created gaps in context-specific validation for the other districts.
- We derived the data for diseases manually due to unavailability of prior datasets. This may introduce a potential for human error, though we verified it by doing the calculations on both MS Excel and Python code.
- For Arsenic testing, the previous researches – on whose research this paper is based – had a limited sample size, and hence the probabilities generated for Arsenic may not be as accurate as other nodes in the model.

VII. Conclusion

Prediction using Bayesian Network models is a highly effective tool, yet it is not widely recognized or implemented. While machine learning approaches are more prevalent today due to the ease of data feeding and learning to carry out similar tasks, they lack a crucial element: the 'why.' This is where the causality aspect of BN models becomes significant. This study may be the first in Pakistan to locally predict water quality and disease outbreaks in Sindh. The data from the STRP study was fed into our BN model, along with probabilities of diseases given certain parameters — inferred and calculated from existing literature — and the model predicted various risk assessments across different districts. Based on the model's results, we recommend immediate safety measures to evaluate and improve water quality in rural areas. These measures can include simple and cost-efficient water treatment methods, such as boiling water before consumption.

Acknowledgment

This research-based model and paper were completed as a final project for a course titled *Probabilistic Graphical Models*, offered by Dr. Saleha Raza at Habib University. Our domain experts, including local and district-level advisors, as well as other contacts in the Public Health sector, provided significant assistance. Moreover, the research conducted by Dr. Humaira Qureshi on local areas of Sindh served as a major motivating factor for our work.

References

[1] J. Ahmed, et al., “Quantitative Microbial Risk Assessment of Drinking Water Quality to Predict the Risk of Waterborne Diseases in Primary-School Children,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 8, pp. xx-yy, Apr. 2020, doi: https://doi.org/10.3390/ijerph17082774.

[2] A. Nayan, J. Saha, A. Mozumder, K. Mahmud, A. Azad, and M. Kibria, “A machine learning approach for early detection of fish diseases by analyzing water quality,” *Trends in Sciences*, vol. 18, no. 21, p. 351, 2021, doi: https://doi.org/10.48048/tis.2021.351.

[3] D. H. Hall and Q.-T. Le, “Use of Bayesian networks in predicting contamination of drinking water with E. coli in rural Vietnam,” *Transactions of The Royal Society of Tropical Medicine and Hygiene*, vol. 111, no. 6, pp. 270–277, Jun. 2017, doi: https://doi.org/10.1093/trstmh/trx043.

[4] A. Ali, S. Javed, S. Ullah, S. H. Fatima, F. Zaidi, and M. S. Khan, “Bayesian spatial analysis and prediction of groundwater contamination in Jhelum city (Pakistan),” *Environmental Earth Sciences*, vol. 77, no. 3, Jan. 2018, doi: https://doi.org/10.1007/s12665-018-7253-5.

[5] A. M. Nangraj, E. Fatima, U. Aftab, H. Qureshi, “Assessment of Water Quality of Hand Pumps from Lower, Middle and Upper Districts of Sindh, Pakistan.”

Appendix

A. Code for Probability Computation

Below is the code used for Calculations of Probabilities; we had a similar approach for the other two districts too, only changing the file path to the excel sheet of respective districts.

```
import pandas as pd

# Load the Excel file to inspect the data structure
file_path = ’D:\\MARIA\\PGM\\project\\Analysis\\sanghar
water parameter results.xlsx’
data = pd.read_excel(file_path)

print("Data Preview:")
print(data.head())

thresholds = {
    ’Hardness mg/L’: 100,
    ’Nitrate mg/L’: 10,
    ’Nitrite mg/L’: 1,
    ’Iron mg/L’: 10,
    ’Chromium/Cr (VI) mg/L’: 2,
    ’Lead ppb’: 15,
    ’Copper mg/L’: 1,
    ’Mercury mg/L’: 0.002,
    ’Fluoride mg/L’: 4,
    ’Coliforms MPN’: 1,
    ’E.Coli MPN’: 1,
}

# Create a new DataFrame to store binary classifications
binary_data = data.copy()

# Apply thresholds to classify parameters
for param, limit in thresholds.items():
    if param in binary_data.columns:
        binary_data[param] = binary_data[param].
        apply(lambda x: 1 if x > limit else (0 if x <= limit
        else None))

# Drop rows with missing data for simplicity
binary_data_clean = binary_data.dropna()


total_entries = len(binary_data_clean)
# Total rows after cleaning

# Calculate probabilities without smoothing
conditional_probabilities = {}
for param in thresholds.keys():
    if param in binary_data_clean.columns:
        num_impermissible =
        binary_data_clean[binary_data_clean[param] == 1].
        shape[0]
        conditional_probabilities[param] =
        num_impermissible / total_entries

# Apply Laplace smoothing to the probabilities
k = 1  # Smoothing factor
smoothed_probabilities = {}
for param in thresholds.keys():
    if param in binary_data_clean.columns:
        num_impermissible =
        binary_data_clean[binary_data_clean[param] == 1].
        shape[0]
        smoothed_probabilities[param] =
        (num_impermissible + k)/(total_entries + 2 * k)

# Print the smoothed probabilities
print("\nConditional Probabilities Laplace Smoothing):")
for param, prob in smoothed_probabilities.items():
    print(f"{param}: {prob:.4f}")
```