Predicting Occurance of Kidney Stones: Exploratory Data Analysis of Relevant Features in NHANES 2017 - 2020

David Nguyen 300584723 (Individual)

Github

Group members
Katie Chu-Fong 300601174; Thinh Nguyen 300611544;
Brook Thomson 300653729

2024-10-19

Executive Summary

Kidney stones affect approximately 10% of people in their lifetime, causing severe pain and potentially leading to serious complications. Our research aimed to develop a predictive model to identify individuals at high risk for kidney stone formation, using data from the National Health and Nutrition Examination Survey. We analyzed various factors and found that several key elements are strongly associated with an increased risk of kidney stones: High blood pressure, Gender (being female), Saturated fatty acid intake (particularly dodecanoic acid), Abdominal pain history, Vitamin A intake, Age, Race (particularly non-Hispanic White and non-Hispanic Black), History of gallstones

Our predictive model demonstrated strong performance in identifying individuals at risk for kidney stones. Importantly, the model revealed that risk factors can vary significantly between individuals. For example, high blood pressure emerged as the most significant risk factor, while certain dietary elements like saturated fatty acid intake also played a crucial role.

These findings highlight the complex nature of kidney stone formation and underscore the importance of personalized risk assessment. By leveraging these insights, healthcare providers can implement targeted preventive measures, potentially reducing the incidence of kidney stones, improving patient outcomes, and decreasing healthcare costs associated with treatment and complications.

Additionally, we recommend implementing targeted screening programs and patient education focusing on modifiable risk factors, especially blood pressure management and personalized dietary recommendations. While our model is promising, further research is needed to fully understand kidney stone formation mechanisms and refine predictive models, particularly regarding the interplay between physiological conditions and dietary factors across diverse populations.

Contents

1. Background

Is kidney stone prevalence associated with factors such as diet, lifestyle, and other existing medical conditions? Kidney stones are solid deposits of minerals and salts that develop in the urinary tract, and can cause blockage and severe pain when urine is passed (National Institutes of Health, n.d.). It is a common condition that affects approximately 10% of people at least once in their lifetime, and in some cases may require significant and/or recurrent treatment (Abufaraj et al., 2020).

This report presents an exploratory data analysis, investigating variables previously shown to be associated with kidney stone occurance. It is based on the National Health and Nutrition Examination Survey (NHANES) from the National Center for Health Statistics, of the Centers for Disease Control and Prevention. NHANES is an ongoing program of surveys in the United States that assesses the health and nutritional status of adults and children. The surveys collect health-related data ranging over a number of topics, which are organised broadly into Demographics, Dietary, Examination, Laboratory, and Questionnaire. It is widely used to analyse or identify associative or causal factors of various conditions and/or diseases, such as diabetes, hypertension, and hearing loss. Thus, the comprehensive scope of NHANES makes it an ideal resource to investigate factors associated with kidney stones, especially given the diverse range of known causes.

In the following sections, we evaluate the completeness, quality, and distributions of kidney stone-relevant data in NHANES. Data from the most recent cycle is used, NHANES 2017 - March 2020.

2. Data Description

2.1 Data Structures and Types

Data from each NHANES cycle is released as many tables, each containing a collection of similar features. For the specific focus on kidney stone disease, only a subset of tables is used, and from these tables, only a subset of key features. The integrated dataset used in this project is composed of 9208 instances/rows, and 146 columns. The column `SEQN` contains a unique identifier for each instance, and the column `KIQ026` contains the target variable. Thus, there are 144 informative features.

The target variable belongs to the Questionnaire component of NHANES, and is phrased as "Ever had kidney stones?". Possible answers of this question are "Yes", "No", "Refused", and "Don't know". Only Yes/No are used as the binary classification label of this project.

Counts (and proportions) of the binary target variable classes are as follows:

- Yes, has had kidney stones: 866 instances (0.09405)
- No, has not had kidney stones: 8342 instances (0.906)

The key features are broadly described in the following:

- **Demographic**: gender, age, race, education, marital status, and income. Men and older individuals are more likely to have had kidney stones (Lieske et al., 2014), and there is evidence that kidney stone prevalence and severity is associated with various socioeconomic factors (Bayne et al., 2019; Trivedi et al., 2008; Winitzki et al., 2022).
- **Dietary**: vitamin, water, nutrient, and dietary supplement intake. Kidney stone incidence increases with certain dietary habits, such as low calcium, low potassium, and low fluid diets (Cappucio et al., 1990; Chewcharat et al., 2022). Everyday foods in the NHANES dietary interviews are deconstructed and aggregated into their nutritional components, thus there is highly specific (and largely correlated) dietary and nutrient data that constitutes a significant portion of the total features explored.
- **Examination**: body mass index (BMI), blood pressure, and pulse readings. Indicators of general health are useful predictive features for kidney stone risk (Cappucio et al., 1990; Jeong et al., 2011).

- **Laboratory**: aspects of biochemistry profile, and urine-associated tests. Kidney diseases or urinary tract abnormalities (that can lead to kidney stones) have been associated with abnormal urinary levels of components such as glucose, lead, and albumin creatinine ratio (Spatola et al., 2016). Studies using NHANES, by Chewcharat & Curhan (2020) and Sun et al. (2019), have previously assessed prevalence of kidney stones in comparison to relevant laboratory values.
- **Questionnaire**: past medical history (conditions and medicines), dietary and alcohol habits, urinary tract function, physical activity, smoking, and sleep habits. Again, general health, behaviours, and lifestyle have a large influence on kidney stone disease (Li et al., 2014; Xue et al., 2024). Factors such as lack of physical activity and smoking can indirectly damage the urinary tract and promote stone formation (Lieske, 2013; Soueidan et al., 2015).

Feature type ranges from numerical continuous and discrete to categorical binary, nominal, and ordinal. Dietary, examination, and laboratory data are mainly numerical, while demographic and questionnaire data are mainly categorical. To avoid difficult or complicated natural languange processing or text mining, free-text data was not selected.

Counts of feature types and brief examples are as follows:

- **97 numerical features**, e.g. energy in kilocalories (continuous); age in years (discrete)
- **49 categorical features**, e.g. gender (binary: male, female); race (nominal: Mexican American, other Hispanic, white, etc.); diet healthiness (ordinal: excellent, very good, fair, etc.)

2.2 Summary of the Data

27 features have no missing values (not including the unique identifier and target variable columns).

Features that do have missing data can be summarised as follows:

- 98 features have under 25% missing data;
- 5 features have 25 - 50% missing data;
- 7 features have 50 - 75% missing data;
- 6 features have 75% - 100% missing data.

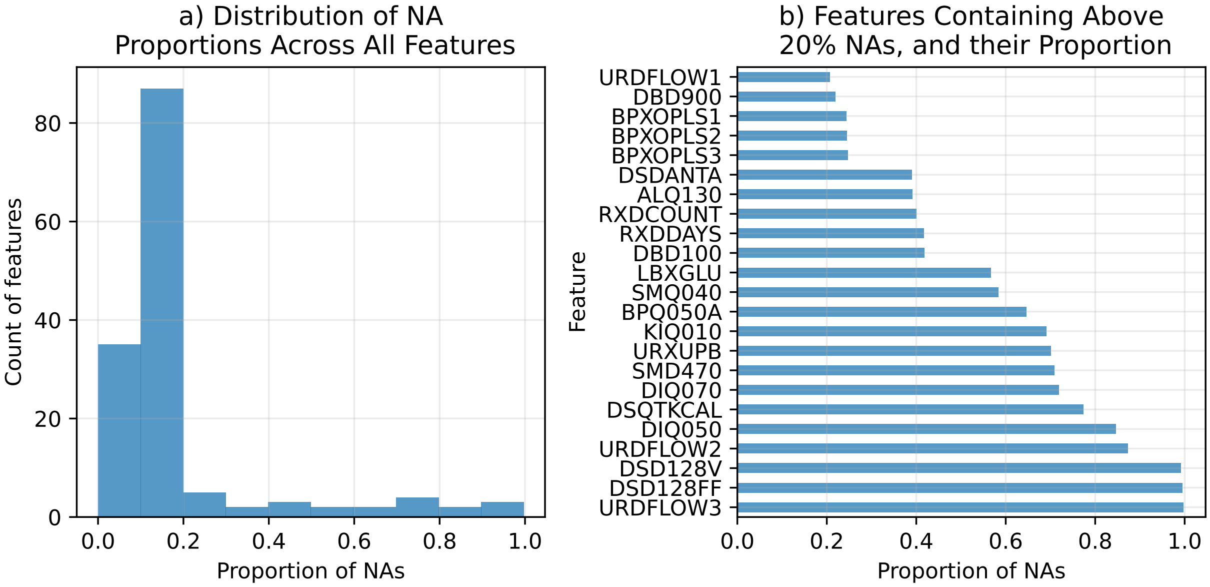


Figure 1: a) Count Of Features With Various Missing Data Proportions. b) Missing Data Proportions Of The Features Containing Above 20% Missing Data.

Overall, the majority of features do not contain a substantial proportion of missing data (Figure 1a).

Over half of features have less than 20% missing data (Figure 1a). Features from the table "Dietary Interview - Total Nutrient Intakes" (P_DR1TOT) are the main contributors to this particular proportion. A large number of features were selected from that table, and data collected within pertains to a consistent subset of people. Consequently, it is reasonable to assume that features originating from the same or similar NHANES tables will share comparable patterns of missing data. For example, features related to dietary intake will only have recorded values for those who partook in dietary interviews, which may differ from the set of people who partook in laboratory tests. This pattern can also be seen in Figure 1b with the set of features `BPXOPLS1`, `BPXOPLS2`, and `BPXOPLS3`, which correspond to successive pulse readings and have identical missing value proportions (~25%).

`KIQ010`, `DSD128V`, and `DSD128FF` are features with very high percentages of missing data (over 50%), as seen in Figure 1b. They have structural missingness - e.g. in the case of `KIQ010`, recording a value for the amount of urine lost is dependent on the participant affirming that they have had urinary leakage, which most participants have not. Likewise, `DSD128V` and `DSD128FF` are both dependent on the participant affiriming that they do take supplements, which, again, may not be the case for most.

3. Ethics, Privacy, and Security

3.1 Ethical Considerations

"…ethical issues can be (a) epistemic, related to misguided, inconclusive or inscrutable evidence; (b) normative, related to unfair outcomes and transformative effectives; or (c) related to traceability" (Morley et al., 2020)

These are the broad scopes of the ethical issues that are present when using machine learning methods in healthcare.

- Epistemic – evidence related ethical issues. This includes bias within a dataset, opacity in decision making, and inaccurate decisions made from inconclusive evidence.
- Normative – learning and transforming processes of algorithms which may result in discriminatory outcomes and/or outcomes where the reasoning is unclear and therefore difficult for a patient to understand, or refute.
- Traceability – the ability to trace mistakes, and identify responsibility for mistakes. Additionally, that information about the algorithm be accessible and comprehensible (Mittelstadt et al., 2016)

These ideas are heavily intertwined, for example, bias in a dataset (epistemic) may result in a discriminatory outcome (normative). While we always want our algorithms to have accurate, unbiased outcomes, the stakes become high with healthcare, as an incorrect result could be fatal to the patient. In the context of kidney stone prediction, it is less a diagnostic tool for current kidney stones, but whether they may develop them in future, so preventative measures could be taken to decrease a person's risk, such as a change in diet, medication or regularity of checkups. While there are not life stakes, there is also the stake of the cost to the patient, and their ability to give informed consent about the medical treatments they recieve – which is difficult to provide if you do not know why you have been deemed at risk. Steps should be taken to make sure the algorithm is traceable, and attempts should be made to reduce any bias prevalent in the dataset.

3.2 Privacy Concerns

Personal health data is extremely sensitive information, making privacy a huge concern. This is a widely recognised concern however, which means there is a great deal of legislation we and NHANES need to be in line with.

As our data is from the NHANES dataset, which is part of a program by a governmental agency of the United Stated of America, and the data is about those who reside in the United States, we must consider the laws, protections, and regulations the USA has, as well as the laws of our own country and the laws of social acceptance. While we may not be the ones who collected the data or released it, we are not relinquished of our moral duties. Using data that doesn't meet privacy standards could lead to legal trouble and societal backlash, so ensuring our data already meets privacy standards - and if not, taking steps to ensure it does - is key.

The United States of America currently has 4 federal laws that cover patient privacy in regards to the NHANES survey data – these are:

1. The Privacy Act of 1974 - requires that data about an individual cannot be released without the individual's consent (U.S Department of Justice, 2014).
2. The Confidential Information Protection and Statistical Efficiency Act – states if information is gathered for statistical purposes, it should only be used for statistical purposes, that an individual should not identifiable, and that confidentiality should be safeguarded by controlling access to the information (Federal Chief Information Officers, 2002).
3. The Cybersecurity Enhancement Act – requires agencies to have appropriate cybersecurity measures and plans in place (Library of Congress, 2014).
4. Section 308(d) of the Public Health Service Act – similar to The Confidential Information Protection and Statistical Efficiency Act, it states that a person should not be identifiable, and that the data may only be used for the purpose it was supplied for (National Center for Health Statistics, 2021; Office of the Legislative Council, 2020).

NHANES is currently in compliance with the above acts, and by using the data for statistical purposes, we are in compliance with The Confidential Information Protection and Statistical Efficiency Act. However, the United States of America also has differing data privacy laws state-by-state (Pittman, 2023), and as the NHANES survey is national, participants will be from a variety of states and therefore have different privacy standards. These state level data privacy laws are typically surrounding consumer data and corporations, as opposed to governmental survey data, but are still important to consider. While geographic information is collected as part of NHANES, its use is limited and restricted, so we are unable to verify ourselves if NHANES is complying with the state laws, and the NHANES privacy page and brochure were both released prior to state privacy laws coming into effect. States are also continuing to roll out new privacy laws and not all are yet in effect, and so even if the data is currently legally sound in terms of privacy, this may change in the coming years. While we may not know NHANES and our own exact legality stance in the states currently, we do know that the information that NHANES releases to the public has no identifiable information, only those with special permissions can access some identifiers (such as geographic location), and these steps ensure patient confidentiality.

3.3 Security Measures

To ensure the security of the NHANES dataset and compliance with the National Center for Health Statistics (NCHS) requirements, our project has implemented a set of data security measures. The NHANES dataset is stored in an access-controlled repository, with raw and processed data kept in separate, restricted directories. Access to these directories is limited to authorized group members only, utilizing robust permission settings to control access based on specific tasks and responsibilities.

All data transfers are conducted using secure, encrypted protocols. We maintain detailed audit trails and leverage version control capabilities to track all data access and modifications. Access permissions are regularly reviewed and updated to align with group members' current tasks and responsibilities, ensuring that access rights remain current and appropriate throughout the project lifecycle.

We have implemented a data minimization strategy, ensuring that only essential NHANES data relevant to the project is stored and processed. Additional security measures include two-factor authentication for all team members accessing the data and encryption for any locally stored data.

Upon completion of the project, all data files and any derived data will be securely deleted from all storage locations using methods that prevent recovery. This ensures that the data is not retained unnecessarily and reduces the risk of unauthorized access or disclosure after the project's conclusion. All these practices are in full compliance with NCHS requirements for handling NHANES data.

4. Exploratory Data Analysis

4.1 Demographic Analysis

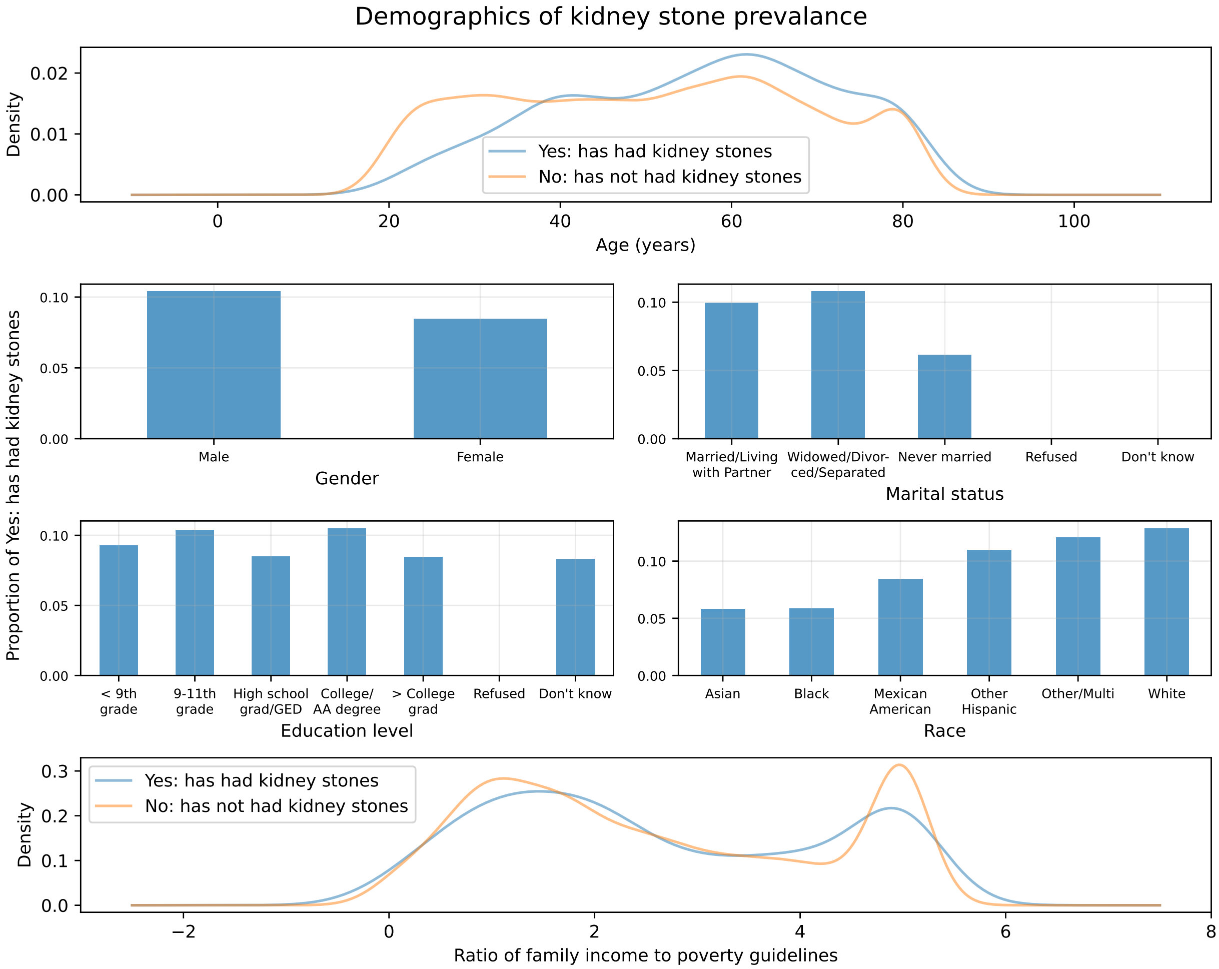


Figure 2: A) Density Distribution Of Age Across Those Who Have Had Kidney Stones And Those Who Have Not. B) Proportion Of Males And Females Within Their Respective Groups Who Have Had Kidney Stones. C) Proportion Of Various Races Within Their Respective Groups Who Have Had Kidney Stones. D) Proportion Of Various Education Levels Within Their Respective Groups Who Have Had Kidney Stones. The Category "Refused" Was Removed Due To Low Count. E) Proportion Of Various Marital Statuses Within Their Respective Groups Who Have Had Kidney Stones. The Categories "Refused" And "Don't Know" Were Removed Due To Low Count. F) Density Distribution Of Ratio Of Family Income To Poverty Guidelines.

At a younger age (20 - 40 years of age), a notably higher proportion of people have never had kidney stones as opposed to have (Figure 2a). As age increases (> 50 years of age), the proportion of people who have never had kidney stones becomes less than those who have. The overall prevalence of having had kidney stones increases steadily from 20 - 40 years of age, plateaus after 40 years of age, then increases again to peak at ~60 years of age.

Figure 2b shows that approximately 10% of males have had kidney stones, while a lesser percentage of around 8% of females have. Thus, kidney stones among males are slightly higher than the overall prevalence of kidney stones (~9.4%), while females are slightly below.

There is clear fluctuation in kidney stone prevalence among different races (Figure 2c), with Asian and Black people at the lowest (just above 5% have had kidney stones), increasing to Mexican Americans (approximately 8%). Races with kidney stone prevalence greater than the overall prevalence are other Hispanic (over 10%), other/multiracial, and White people (latter two are close to 15%). There is a large distinction (~10%) between the lowest and highest prevalence. The low and high proportions are also significantly different from the overall kidney stone prevalence.

As education level changes, the proportion of those who have had kidney stones fluctuates, but there is no obvious trend among successive education levels (Figure 2d). The difference between the education level with the highest kidney stone prevalence (college/AA degree at ~10%) and lowest (high school grad/GED at ~8%) is relatively minimal.

Those that are married/living with partner or widowed/divorced/separated show a greater prevalence of kidney stones than those that have never married, as seen in Figure 2e. Never married people also have a much lower prevalence than overall kidney stone prevalance. However, this may be due to the confounding factor of age, instead of an inherent characteristic of marriage that increases kidney stone occurance.

In Figure 2f, a ratio close to 1 means family income is approximately equal to poverty thresholds; greater than 1 means family income is above poverty thresholds. At a lower to middle ratio (0 - 4), it is slightly more common to not have had kidney stones, but not significantly. At a higher ratio (> 4), there is a large difference - the prevalence of never having had kidney stones is notably higher than having had them.

Overall, Figure 2 indicates that nearly all demographic features - age, gender, marital status, race, and ratio of family income to poverty guidelines - are associated with prevalence of kidney stones. Confirming previous research, older people and males are more likely to have had kidney stones. Age may be a confounding factor in the apparent association of kidney stone occurance with marital status, but regardless it can still be a useful feature. The lack of significant association between education level and kidney stones indicate that it might be uninformative in a predictive context.

4.2 Health Conditions Analysis

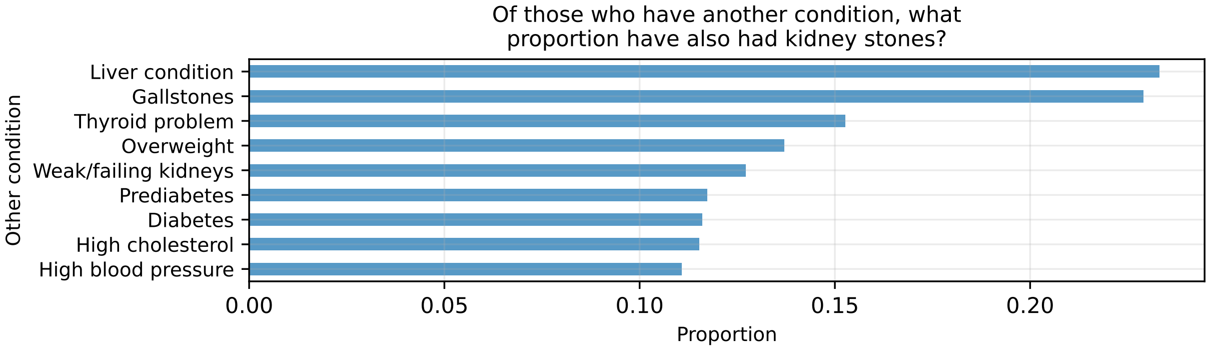


Figure 3: The Proportion Of People With Various Other Conditions Who Also Have Had Kidney Stones.

Gallstones and weak/failing kidneys are the most strongly associated with kidney stone occurance, with close to 25% of people having (either or both of) those conditions also having had kidney stones. Between ~11% and ~15% of people who have the remaining conditions also have had kidney stones. All these are much higher than overall kidney stone prevalence (9.4%), indicating that these features are likely to be useful for a predictive model. It can be noted that some of these conditions may also possess a high degree of correlation between each other, which may be reflected in their similar proportions in Figure 3, e.g. being overweight and having high cholesterol. Feature combination/transformation could be used to reduce dimensionality.

4.3 Laboratory Analysis

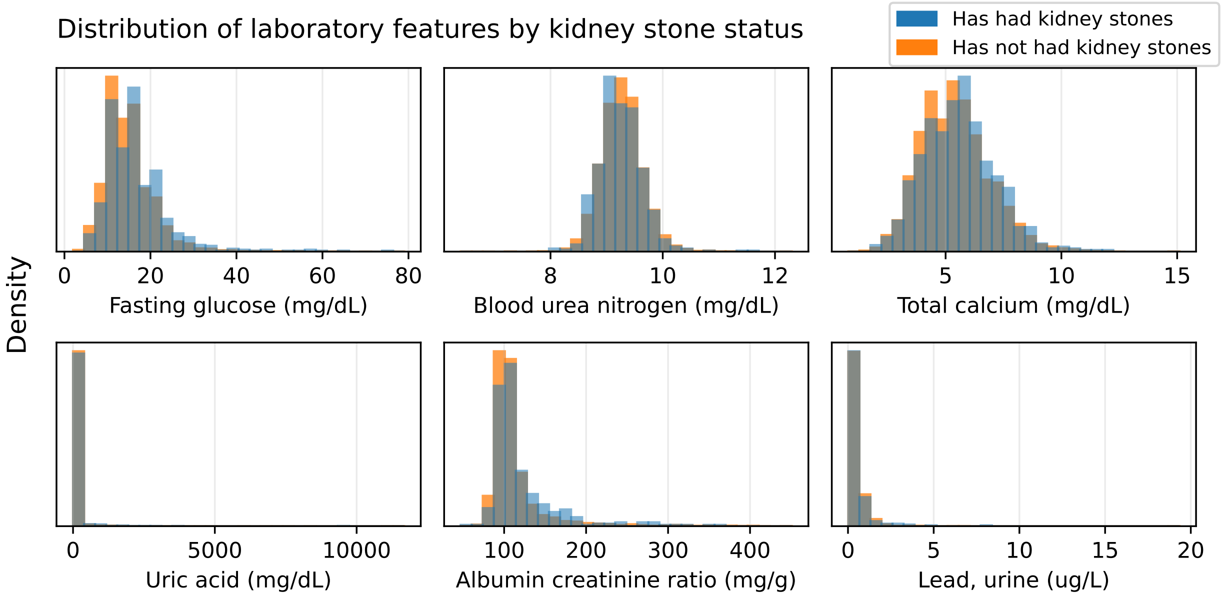


Figure 4: Density Distributions Of Laboratory Features, Split By Those Who Have Had Kidney Stones And Have Not Had Kidney Stones. A) Fasting Glucose. B) Blood Urea Nitrogen. C) Total Calcium. D) Uric Acid. E) Albumin Creatinine Ratio. F) Lead In Urine

Figure 4 shows that most laboratory features appear within expected ranges, with the exception of uric acid (Figure 4d). There appears to be outlier(s) skewing this feature with up to 10000 mg/dL uric acid, which is likely to be an error as ordinary uric acid levels should not exceed the single-digit mg/dL range.

Distribution shape of laboratory features remains relatively identical, regardless of kidney stone status. Distributions for fasting glucose (Figure 4a) and total calcium (Figure 4c) are shifted slightly right (towards higher values) for those who have had kidney stones. The peak bin for blood urea nitrogen (Figure 4b) is at a marginally lower value for those who have had kidney stones in comparison to those who have not. Albumin creatinine ratio appears to peak later, and remain slightly higher, at increasing mg/g for those who have had kidney stones (Figure 4e). Distributions for lead (Figure 4f) and uric acid are consistent for both kidney stone statuses - however, detail in the uric acid histogram may be obscured by the outlier(s).

Therefore this indicates that uric acid and lead are not associated with kidney stone prevalence. The remaining laboratory features are associated due to differing distributions according to kidney stone status.

4.4 Dietary Analysis

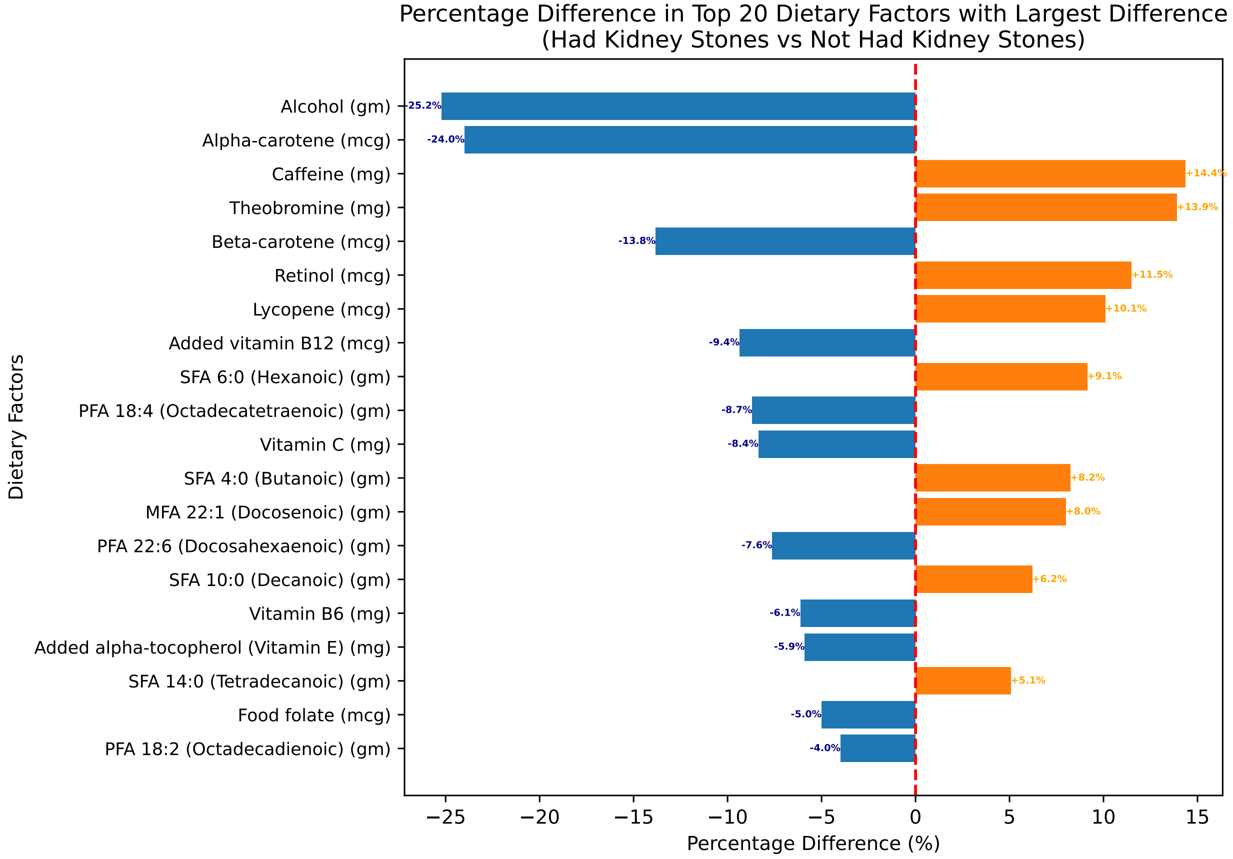


Figure 5: Dietary Differences.

Alcohol consumption shows the largest difference, with individuals who have had kidney stones consuming 25.2% less alcohol. This aligns with studies suggesting that excessive alcohol intake can increase kidney stone risk by causing dehydration and altering urine composition (Ferraro et al., 2013).

In addition, alpha-carotene and beta-carotene (both forms of vitamin A) display substantial negative differences (-24.0% and -13.8% respectively), with those who have had kidney stones consuming less. In contrast, other form of vitamin A, Retinol intake is 11.5% higher in the kidney stone group, contrasting with the carotenoid findings. Some research has indicated that excessive vitamin A intake may increase kidney stone risk (Tang et al., 2012), which could explain the lower carotenoid but higher retinol intake in those with a history of stones.

Caffeine and theobromine consumption is notably higher in those with a history of kidney stones (14.4% and 13.9% more, respectively). While caffeine has been associated with increased risk of kidney stones in some studies (Ferraro et al., 2014), the higher intake in those with a history of stones could reflect changes in fluid consumption patterns post-diagnosis.

Vitamin C intake is 8.4% lower in individuals who have had kidney stones. This aligns with studies suggesting that high-dose vitamin C supplementation may increase kidney stone risk (Thomas et al., 2013), potentially leading to reduced intake in those with a history of stones.

Among the top factors, we see a trend in vitamins and antioxidants, particularly forms of vitamin A, vitamin C, and vitamin E (alpha-tocopherol).

4. Detailed Analysis Results

4.1 Modelling Methodology Overview

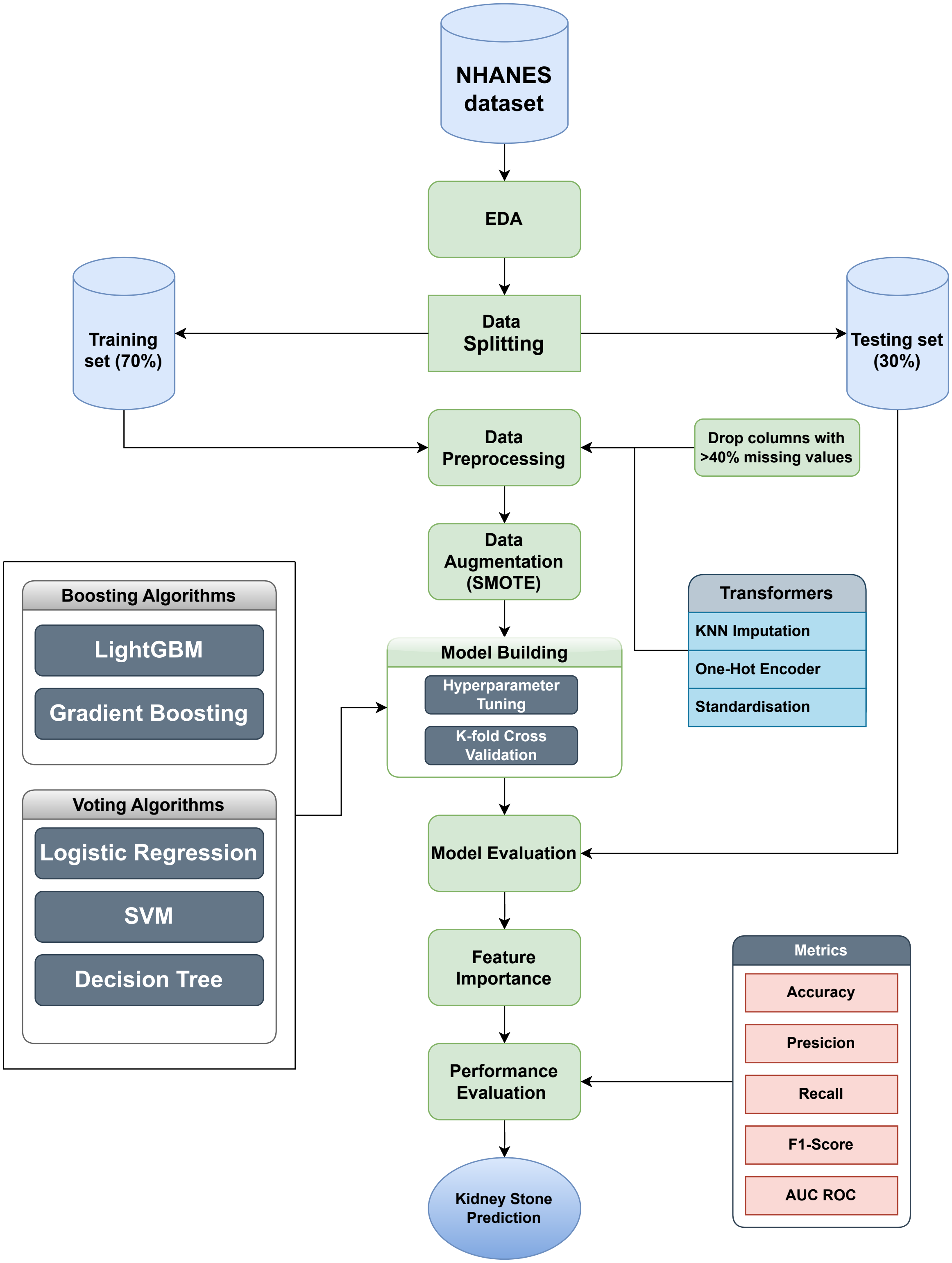


Figure 6: Overview Of The Kidney Stones Disease Prediction Methodology.

4.2 Data Preprocessing and Data Augmentation

4.2.1 Data Preprocessing Techniques

Our data preprocessing pipeline comprised several key steps to prepare the dataset for model training:

Dataset Splitting: We initially partitioned the dataset into training and testing sets using a 70-30 ratio. This split ensures a substantial portion of data for model training while reserving a significant subset for unbiased evaluation.

Feature Selection: To improve data quality and model efficiency, we removed features with more than 40% missing values. This step reduced the number of features from 145 to 128, focusing our analysis on the most complete and potentially informative variables.

Missing Value Imputation: For the remaining features, both numeric and categorical, we employed K-Nearest Neighbors (KNN) imputation with k=5 neighbors. This method estimates missing values based on similar instances, potentially preserving important data relationships.

Feature Scaling: Numeric features were standardized using Standard Scaler, transforming them to have zero mean and unit variance.

Categorical Encoding: Categorical variables were encoded using one-hot encoding, with the 'ignore' option for handling unknown categories during model inference.

4.2.2 Data Augmentation

In our initial training set, we observed a significant class imbalance, with 6,445 total records comprising 5,818 patients without kidney stones (90.3%) and only 627 patients with kidney stones (9.7%). This imbalance posed a potential challenge, as models could be biased towards the majority class, potentially leading to poor predictive performance for the minority class.

To mitigate this issue, we employed the **Synthetic Minority Over-sampling Technique (SMOTE)**. SMOTE works by creating synthetic examples of the minority class, effectively increasing its representation in the dataset without simply duplicating existing instances.

Post-SMOTE, our training set achieved balance with 5,818 records in each class, resulting in a total of 11,636 records. With equal representation of both classes, our models can learn patterns associated with kidney stone presence more effectively, and the risk of model bias towards the majority class is significantly diminished.

For a visual representation of the data distribution before and after SMOTE balancing, please refer to Appendix A.

4.3 Model Development

4.3.1 Model Selection

For this study, we selected three distinct classification models, each representing different algorithmic approaches:

Gradient Boosting Classifier: An ensemble method that builds a series of weak learners (typically decision trees) sequentially, with each new model correcting errors from the previous ones.

LightGBM Classifier: A gradient boosting framework that uses tree-based learning algorithms, known for its efficiency and ability to handle large-scale data.

Voting Classifier: An ensemble model that combines Logistic Regression, Decision Tree, and Support Vector Machine (SVM) classifiers. This approach leverages the strengths of different algorithms to make predictions.

4.3.2 Classifier Performance

Each model was trained using its optimal hyperparameters on the SMOTE-augmented training data. This approach ensures that the models are well-tuned and robust to class imbalance.

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
GradientBoosting	0.9102	0.8523	0.9102	0.8726	0.6313
LightGBM	0.9121	0.8565	0.9121	0.8729	0.6282
VotingClassifier	0.8625	0.8494	0.8625	0.8558	0.5858

Table 1: Performance Comparison Between Classifiers.

All three models demonstrate high accuracy, with scores above 0.86, indicating strong overall predictive capability for kidney stone occurrence. However, note that it is not a very balanced dataset, so the accuracy alone is not the good metric to consider. The ROC AUC scores are lower than might be expected given the high accuracy scores, ranging from 0.5858 to 0.6313. This discrepancy could indicate that while the models are good at classifying the majority of cases, they might struggle with some harder-to-distinguish instances.

Specifically, **LightGBM Classifier** slightly outperforms the other models across most metrics. This suggests that **LightGBM** is particularly effective at capturing the underlying patterns in our dataset. **Gradient Boosting** shows very similar performance , with only marginal differences across ROC AUC.

4.4 Interpretation of Model Results

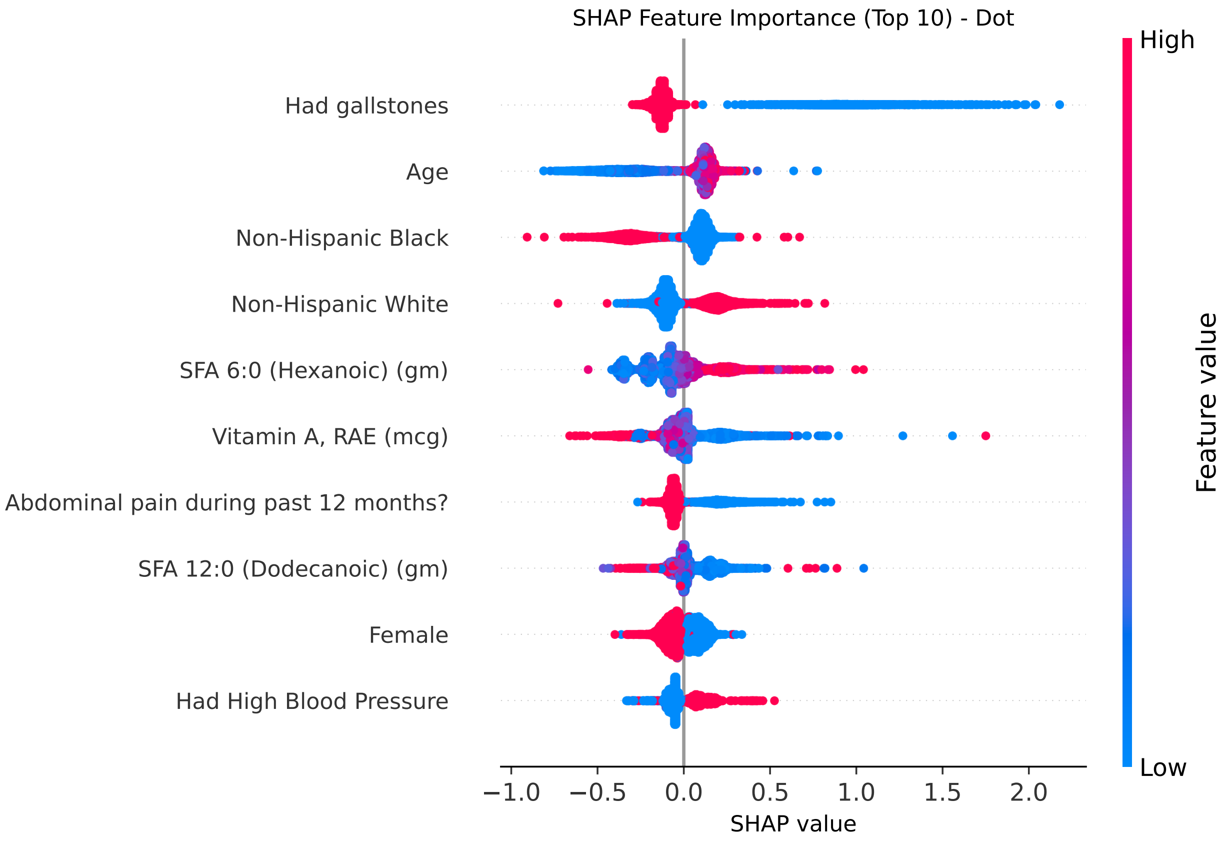


Figure 7: Top 10 Predictors of Kidney Stone Risk: Impact and Distribution (SHAP Summary Plot)

The beeswarm plot visualizes the distribution of SHAP values for each feature across all samples, revealing the magnitude and direction of a feature's impact on model predictions.

For gallstones, the predominantly positive SHAP values (red points) indicate an increased likelihood of kidney stones, though the impact varies across cases. Age demonstrates a positive trend, with higher SHAP values at increased ages, suggesting elevated risk with aging.

Interestingly, Non-Hispanic Black shows largely negative SHAP values, indicating a reduced risk, while Non-Hispanic White displays mainly positive values, suggesting an increased risk.

Among dietary predictors, short-chain saturated fatty acids (SFA 6:0 and 12:0) and vitamin A exhibit wide-ranging impacts, with both positive and negative SHAP values, indicating their effect may be modulated by other factors. Very long-chain polyunsaturated fatty acids (PFA 20:4) lean towards negative values, suggesting a potential protective effect.

Abdominal pain's primarily positive SHAP values underscore its clinical significance in predicting kidney stones. The albumin creatinine ratio's broad SHAP value distribution points to a variable impact, likely influenced by individual patient characteristics.

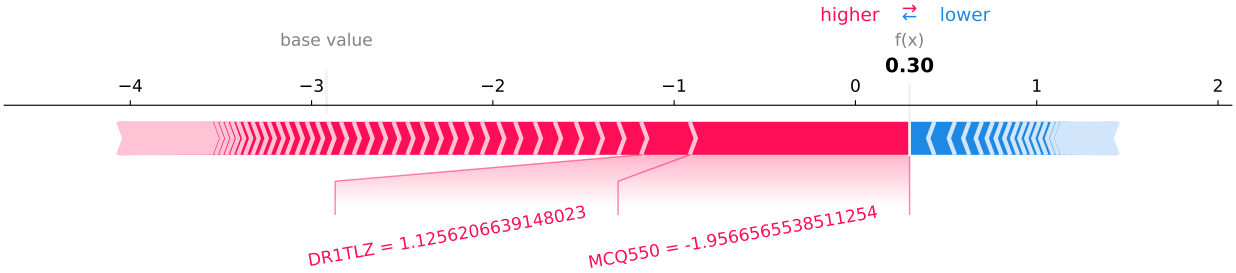


Figure 8: Individual Risk Factor Analysis for Kidney Stone Formation (SHAP Force Plot)

This SHAP force plot illustrates the impact of two variables on the model's prediction for a specific instance, with the final prediction f(x) of 0.30 being greater than the base value. This positive difference indicates that for this particular case, the combined effect of the variables increases the prediction above the average model output. The plot highlights MCQ550 (history of gallstones) with a substantial negative impact of -1.9567, suggesting that a history of gallstones significantly lowers the prediction. Conversely, DR1TLZ (Lutein + Zeaxanthin intake in mcg) shows a positive influence of 1.1256, partially offsetting the negative effect of gallstones. The interplay between these factors results in a net positive effect, pushing the final prediction above the base value. This suggests that while a history of gallstones tends to decrease the predicted outcome, higher intake of Lutein and Zeaxanthin mitigates this effect to some extent, ultimately leading to a prediction that is higher than the model's average output for the population.

5. Conclusions and Recommendations

5.1 Conclusions

Our kidney stone prediction model successfully identified several key factors associated with increased risk of kidney stone formation. These include age, race (particularly Non-Hispanic White), presence of certain health conditions (high blood pressure, gallstones), body mass index, and specific biochemical markers (albumin-creatinine ratio).

Dietary factors, including intake of certain fatty acids (e.g., SFA 6:0 Hexanoic) and vitamins (particularly Vitamin A), play a complex role in kidney stone risk, with their effects not always being straightforward.

The model also demonstrates the multifactorial nature of kidney stone risk, incorporating demographic, health, dietary, and biochemical factors to provide a comprehensive risk assessment.

5.2 Recommendations

Healthcare providers should consider using multifactorial risk assessment tools, like our model, to identify individuals at high risk of kidney stones. This could enable more targeted preventive strategies. Patient education should emphasize the importance of modifiable risk factors, particularly hydration and dietary choices. The protective effect of adequate fluid intake should be a key message. Healthcare systems should also consider implementing screening programs for kidney stone risk, especially for individuals with identified high-risk factors such as high blood pressure or gallstones.

5.3 Limitations

The model is based on cross-sectional data from the NHANES survey, which limits our ability to establish causal relationships or track risk factors over time. While our model incorporates a wide range of factors, it may not capture all possible influences on kidney stone formation, such as genetic predisposition or certain environmental factors. The model's performance may vary across different populations not well-represented in the NHANES dataset. Some of the relationships identified, such as the impact of specific fatty acids or vitamins, require further research to fully understand their mechanisms of influence on kidney stone formation.

6. References

Abufaraj, M., Xu, T., Cao, C., Waldhoer, T., Seitz, C., D'andrea, D., Siyam, A., Tarawneh, R., Fajkovic, H., Schernhammer, E., Yang, L., & Shariat, S. F. (2020). Prevalence and Trends in Kidney Stone Among Adults in the USA: Analyses of National Health and Nutrition Examination Survey 2007–2018 Data. European Urology Focus, 7(6). https://doi.org/10.1016/j.euf.2020.08.011

Bayne, D. B., Usawachintachit, M., Armas-Phan, M., Tzou, D. T., Wiener, S., Brown, T. T., Stoller, M., & Chi, T. L. (2019). Influence of Socioeconomic Factors on Stone Burden at Presentation to Tertiary Referral Center: Data From the Registry for Stones of the Kidney and Ureter. Urology, 131, 57–63. https://doi.org/10.1016/j.urology.2019.05.009

Cappuccio, F. P., Strazzullo, P., & Mancini, M. (1990). Kidney stones and hypertension: population based study of an independent clinical association. BMJ, 300(6734), 1234–1236. https://doi.org/10.1136/bmj.300.6734.1234

Chen, Y., Lee, J., Shen, J.-T., Wu, Y., Tsao, Y.-H., Jhan, J., Wang, H.-S., Lee, Y., Huang, S.-P., Chen, S.-C., & Geng, J.-H. (2023). The impact of secondhand smoke on the development of kidney stone disease is not inferior to that of smoking: a longitudinal cohort study. BMC Public Health, 23(1). https://doi.org/10.1186/s12889-023-16116-6

Chewcharat, A., & Curhan, G. (2020). Trends in the prevalence of kidney stones in the United States from 2007 to 2016. Urolithiasis, 49, 27–39. https://doi.org/10.1007/s00240-020-01210-w

Chewcharat, A., Thongprayoon, C., Vaughan, L. E., Mehta, R. A., Schulte, P. J., O'Connor, H. M., Lieske, J. C., Taylor, E. N., & Rule, A. D. (2022). Dietary Risk Factors for Incident and Recurrent Symptomatic Kidney Stones. Mayo Clinic Proceedings, 97(8), 1437–1448. https://doi.org/10.1016/j.mayocp.2022.04.016

Coenen, P., Huysmans, M. A., Holtermann, A., Krause, N., van Mechelen, W., Straker, L. M., & van der Beek, A. J. (2018). Do highly physically active workers die early? A systematic review with meta-analysis of data from 193 696 participants. British Journal of Sports Medicine, 52(20), 1320–1326. https://doi.org/10.1136/bjsports-2017-098540

Curhan, G. C. (2007). Epidemiology of Stone Disease. Urologic Clinics of North America, 34(3), 287–293. https://doi.org/10.1016/j.ucl.2007.04.003

Federal Chief Information Officers. (2002). Confidential Information Protection and Statistical Efficiency Act. Federal Chief Information Officers; U.S. Federal Government. https://www.cio.gov/handbook/it-laws/cipsea/

Ferraro, P. M., Curhan, G. C., Sorensen, M. D., Gambaro, G., & Taylor, E. N. (2015). Physical Activity, Energy Intake and the Risk of Incident Kidney Stones. Journal of Urology, 193(3), 864–868. https://doi.org/10.1016/j.juro.2014.09.010

Ferraro, P. M., Taylor, E. N., Gambaro, G., & Curhan, G. C. (2013). Soda and Other Beverages and the Risk of Kidney Stones. Clinical Journal of the American Society of Nephrology, 8(8), 1389–1395. https://doi.org/10.2215/cjn.11661112

Ferraro, P. M., Taylor, E. N., Gambaro, G., & Curhan, G. C. (2014). Caffeine intake and the risk of kidney stones. The American Journal of Clinical Nutrition, 100(6), 1596–1603. https://doi.org/10.3945/ajcn.114.089987

Hellerstein, J. (2008). Quantitative Data Cleaning for Large Databases. https://dsf.berkeley.edu/jmh/papers/cleaning-unece.pdf

Jeong, I. G., Kang, T., Bang, J. K., Park, J., Kim, W., Hwang, S. S., Kim, H. K., & Park, H. K. (2011). Association Between Metabolic Syndrome and the Presence of Kidney Stones in a Screened Population. American Journal of Kidney Diseases, 58(3), 383–388. https://doi.org/10.1053/j.ajkd.2011.03.021

Li, C. - H., Sung, F. - C., Wang, Y. - C., Lin, D., & Kao, C. - H. (2014). Gallstones increase the risk of developing renal stones: a nationwide population-based retrospective cohort study. QJM: An International Journal of Medicine, 107(6), 451–457. https://doi.org/10.1093/qjmed/hcu017

Library of Congress. (2014, December 18). S.1353 - 113th Congress (2013-2014): Cybersecurity Enhancement Act of 2014. Library of Congress. https://www.congress.gov/bill/113th-congress/senate-bill/1353

Lieske, J. C. (2013). New Insights Regarding the Interrelationship of Obesity, Diet, Physical Activity, and Kidney Stones. Journal of the American Society of Nephrology, 25(2), 211–212. https://doi.org/10.1681/asn.2013111189

Lieske, J. C., Rule, A. D., Krambeck, A. E., Williams, J. C., Bergstralh, E. J., Mehta, R. A., & Moyer, T. P. (2014). Stone composition as a function of age and sex. Clinical Journal of the American Society of Nephrology: CJASN, 9(12), 2141–2146. https://doi.org/10.2215/CJN.05660614

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of algorithms: Mapping the Debate. Big Data & Society, 3(2), 1–21.Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. Social Science & Medicine, 260(113172). https://doi.org/10.1016/j.socscimed.2020.113172

National Center for Health Statistics. (2013). Vital and Health Statistics Report Series 1, Number 56 August 2013. National Center for Health Statistics. https://www.cdc.gov/nchs/data/series/sr_01/sr01_056.pdf

National Center for Health Statistics. (2021, July 22). NHANES - Your Privacy Matters. National Center for Health Statistics; Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/participant/participant-confidentiality.htm

National Institutes of Health. (n.d.). Kidney Stones. National Institute of Diabetes and Digestive and Kidney Diseases. https://www.niddk.nih.gov/health-information/urologic-diseases/kidney-stones?dkrd=hispt0421

NHS. (2018, October 3). Kidney stones - Causes. NHS. https://www.nhs.uk/conditions/kidney-stones/causes/

Office of the Legislative Counsel. (2020). Public Health Service Act. Office of the Legislative Counsel. https://www.govinfo.gov/content/pkg/COMPS-8773/pdf/COMPS-8773.pdf

Prezioso, D., Strazzullo, P., Lotti, T., Bianchi, G., Borghi, L., Caione, P., Carini, M., Caudarella, R., Gambaro, G., Gelosa, M., Guttilla, A., Illiano, E., Martino, M., Meschi, T., Messa, P., Miano, R., Napodano, G., Nouvenne, A., Rendina, D., & Rocco, F. (2015). Dietary treatment of urinary risk factors for renal stone formation. A review of CLU Working Group. Archivio Italiano Di Urologia E Andrologia, 87(2), 105. https://doi.org/10.4081/aiua.2015.2.105

Sorensen, M. D., Chi, T., Shara, N. M., Wang, H., Hsi, R. S., Orchard, T., Kahn, A. J., Jackson, R. D., Miller, J., Reiner, A. P., & Stoller, M. L. (2014). Activity, energy intake, obesity, and the risk of incident kidney stones in postmenopausal women: a report from the Women's Health Initiative. Journal of the American Society of Nephrology: JASN, 25(2), 362–369. https://doi.org/10.1681/ASN.2013050548

Sorensen, M. D., Hsi, R. S., Chi, T., Shara, N., Wactawski-Wende, J., Kahn, A. J., Wang, H., Hou, L., & Stoller, M. L. (2014). Dietary Intake of Fiber, Fruit and Vegetables Decreases the Risk of Incident Kidney Stones in Women: A Women's Health Initiative Report. Journal of Urology, 192(6), 1694–1699. https://doi.org/10.1016/j.juro.2014.05.086

Soueidan, M., Bartlett, S. J., Noureldin, Y. A., Andersen, R. E., & Andonian, S. (2015). Leisure time physical activity, smoking and risk of recent symptomatic urolithiasis: Survey of stone clinic patients. Canadian Urological Association Journal, 9(7-8), 257. https://doi.org/10.5489/cuaj.2879

Spatola, L., Angelini, C., Badalamenti, S., Maringhini, S., & Gambaro, G. (2016). Kidney stones diseases and glycaemic statuses: focus on the latest clinical evidences. Urolithiasis, 45(5), 457–460. https://doi.org/10.1007/s00240-016-0956-8

Sun, Y., Zhou, Q., & Zheng, J. (2019). Nephrotoxic metals of cadmium, lead, mercury and arsenic and the odds of kidney stones in adults: An exposure-response analysis of NHANES 2007–2016. Environment International, 132, 105115. https://doi.org/10.1016/j.envint.2019.105115

Tang, J., McFann, K. K., & Chonchol, M. B. (2012). Association between serum 25-hydroxyvitamin D and nephrolithiasis: the National Health and Nutrition Examination Survey III, 1988-94. Nephrology Dialysis Transplantation, 27(12), 4385–4389. https://doi.org/10.1093/ndt/gfs297

Thomas, L. D. K., Elinder, C.-G., Tiselius, H.-G., Wolk, A., & Åkesson, A. (2013). Ascorbic Acid Supplements and Kidney Stone Incidence Among Men: A Prospective Study. JAMA Internal Medicine, 173(5), 386. https://doi.org/10.1001/jamainternmed.2013.2296

Trivedi, R. B., Ayotte, B., Edelman, D., & Bosworth, H. B. (2008). The association of emotional well-being and marital status with treatment adherence among patients with hypertension. Journal of Behavioral Medicine, 31(6), 489–497. https://doi.org/10.1007/s10865-008-9173-4

U.S. Department of Justice. (2014, June 16). Office of Privacy and Civil Liberties | Privacy Act of 1974. Office of Privacy and Civil Liberties; U.S Department of Justice. https://www.justice.gov/opcl/privacy-act-1974#:~:text=The%20Privacy%20Act%20prohibits%20the

University of Florida Health. (2019). Kidney stones. Department of Urology; University of Florida Health. https://ufhealth.org/conditions-and-treatments/kidney-stones

Wang, X., Sun, M., Wang, L., Li, J., Xie, Z., Guo, R., Wang, Y., & Li, B. (2023). The role of dietary inflammatory index and physical activity in depressive symptoms: Results from NHANES 2007–2016. Journal of Affective Disorders, 335, 332–339. https://doi.org/10.1016/j.jad.2023.05.012

Winitzki, D., Zacharias, H. U., Nadal, J., Baid-Agrawal, S., Schaeffner, E., Schmid, M., Busch, M., Bergmann, M. M., Schultheiss, U., Kotsis, F., Stockmann, H., Meiselbach, H., Wolf, G., Krane, V., Sommerer, C., Eckardt, K.-U., Schneider, M. P., Schlieper, G., Floege, J., & Saritas, T. (2022). Educational Attainment Is Associated With Kidney and Cardiovascular Outcomes in the German CKD (GCKD) Cohort. Kidney International Reports, 7(5). https://doi.org/10.1016/j.ekir.2022.02.001

Xue, W., Xue, Z., Liu, Y., Yin, P., Liu, L., Qu, S., Wu, S., & Yang, C. (2024). Is Kidney Stone Associated with Thyroid Disease? The United States National Health and Nutrition Examination Survey 2007–2018. Endocrine, Metabolic & Immune Disorders, 24(11). https://doi.org/10.2174/0118715303268738231129093935

Yencilek, E., Yencilek, F., Ozcan, C., Demirel, A., Coskun, S., & Basaran, M. (2010). The effect of lycopene on the kidney stones: a preliminary study. Urological Research, 38(4).

Appendices

Appendix A: Confusion Matrix Comparison

Table A1: Confusion Matrix comparison of Gradient Boosting, Light GBM and Voting Classifiers.

Model	True Positive	True Negative	False Positive	False Negative
Gradient Boosting	3	2512	12	236
LightGBM	2	2518	6	237
Voting Classifiers	26	2357	167	213

Appendix B: Performance Metrics Formulas

The evaluated model performance using a comprehensive set of metrics:

Accuracy: The proportion of correct predictions (both true positives and true negatives) among the total number of cases examined.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

TP : The number of samples correctly classified as had kidney stones

TN : The number of samples correctly classified as not had kidney stones

FP : The number of samples incorrectly classified as had kidney stones

FN : The number of samples incorrectly classified as not had kidney stones

Precision: The proportion of correct positive identifications.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also known as Sensitivity): The proportion of actual positive cases that were correctly identified.

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN}$$

F1 Score: The harmonic mean of **Precision** and **Recall**, providing a single score that balances both measures.

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

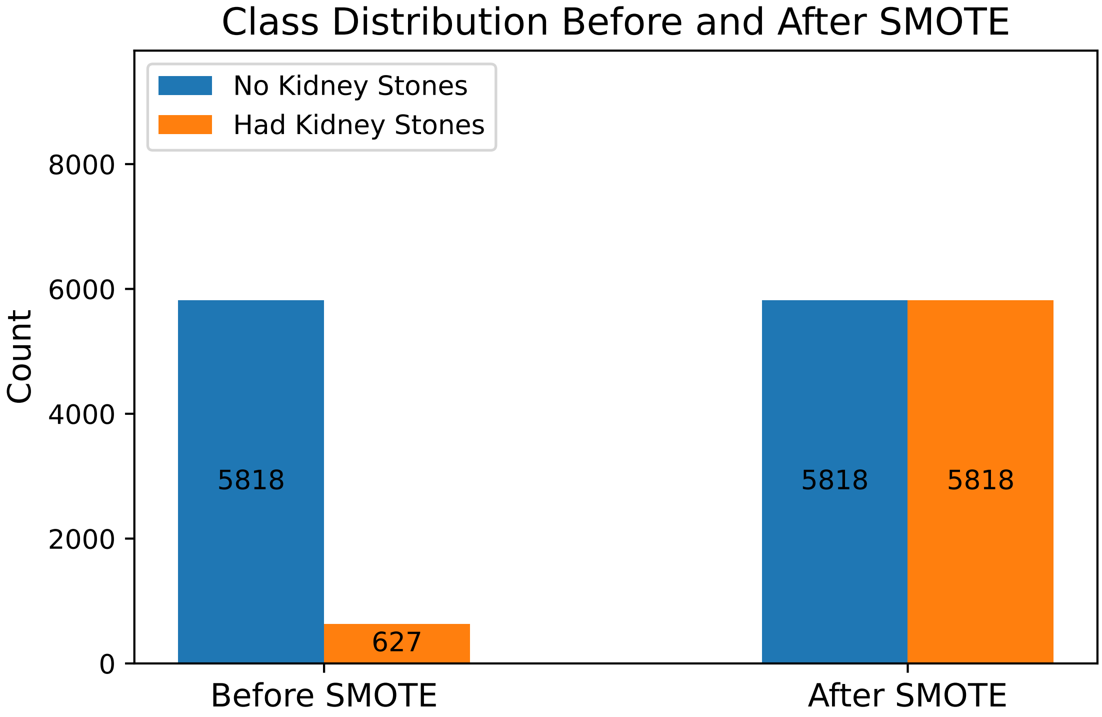
ROC AUC: A measure of the model's ability to distinguish between classes. It's calculated by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings.

$$\text{TPR} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN}$$

The AUC is then calculated as the area under this ROC curve, with values ranging from 0 to 1. An AUC of 0.5 represents a model with no discriminative ability, while an AUC of 1.0 represents a perfect model.

Appendix C: Data Balancing Visualization


Class Distribution Before and After SMOTE
No Kidney Stones
Had Kidney Stones
Count
Before SMOTE
After SMOTE