AI

MDPI

Article

Leveraging Explainable Artificial Intelligence (XAI) for Expert Interpretability in Predicting Rapid Kidney Enlargement Risks in Autosomal Dominant Polycystic Kidney Disease (ADPKD)

Latifa Dwiyanti [1,2,*,†] on behalf of PKDOC, Hidetaka Nambo [1,*] and Nur Hamid [1]

[1] Graduate School of Natural Science and Technology, Kanazawa University, Kanazawa 920-1192, Japan; nur.hamid@stu.kanazawa-u.ac.jp
[2] School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Bandung 40132, Jawa Barat, Indonesia
* Correspondence: latifa@stu.kanazawa-u.ac.jp (L.D.); nambo@blitz.ec.t.kanazawa-u.ac.jp (H.N.)
† Data used in the preparation of this article were obtained from the Polycystic Kidney Disease Outcomes Consortium (PKDOC). As such, the investigators within PKDOC contributed to the design and implementation of the PKDOC database and/or provided data, but did not participate in the analysis of the data or the writing of this report.

Abstract: Autosomal dominant polycystic kidney disease (ADPKD) is the predominant hereditary factor leading to end-stage renal disease (ESRD) worldwide, affecting individuals across all races with a prevalence of 1 in 400 to 1 in 1000. The disease presents significant challenges in management, particularly with limited options for slowing cyst progression, as well as the use of tolvaptan being restricted to high-risk patients due to potential liver injury. However, determining high-risk status typically requires magnetic resonance imaging (MRI) to calculate total kidney volume (TKV), a time-consuming process demanding specialized expertise. Motivated by these challenges, this study proposes alternative methods for high-risk categorization that do not rely on TKV data. Utilizing historical patient data, we aim to predict rapid kidney enlargement in ADPKD patients to support clinical decision-making. We applied seven machine learning algorithms—Random Forest, Logistic Regression, Support Vector Machine (SVM), Light Gradient Boosting Machine (LightGBM), Gradient Boosting Tree, XGBoost, and Deep Neural Network (DNN)—to data from the Polycystic Kidney Disease Outcomes Consortium (PKDOC) database. The XGBoost model, combined with the Synthetic Minority Oversampling Technique (SMOTE), yielded the best performance. We also leveraged explainable artificial intelligence (XAI) techniques, specifically Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP), to visualize and clarify the model's predictions. Furthermore, we generated text summaries to enhance interpretability. To evaluate the effectiveness of our approach, we proposed new metrics to assess explainability and conducted a survey with 27 doctors to compare models with and without XAI techniques. The results indicated that incorporating XAI and textual summaries significantly improved expert explainability and increased confidence in the model's ability to support treatment decisions for ADPKD patients.

Keywords: autosomal dominant polycystic kidney disease (ADPKD); explainable artificial intelligence (XAI); machine learning classification algorithms; user-centered design

Citation: Dwiyanti, L., on behalf of PKDOC; Nambo, H.; Hamid, N. Leveraging Explainable Artificial Intelligence (XAI) for Expert Interpretability in Predicting Rapid Kidney Enlargement Risks in Autosomal Dominant Polycystic Kidney Disease (ADPKD). *AI* **2024**, *5*, 2037–2065. https://doi.org/10.3390/ai5040100

Academic Editors: Rosina O. Weber and Mobyen Uddin Ahmed

Received: 19 August 2024
Revised: 21 October 2024
Accepted: 21 October 2024
Published: 28 October 2024

1. Introduction

Autosomal dominant polycystic kidney disease (ADPKD) is the most common genetic cause of end-stage renal disease (ESRD) around the world. It affects about 133,893 people in 19 European Union (EU) countries, about 500,000 people in the United States (US), and about 31,000 people in Japan [1–3]. ADPKD affects people of all races, with a global prevalence of around 12.5 million [1]. It is defined by the presence of clusters of fluid-filled cysts in both kidneys, leading to a steady deterioration in kidney function. The prevalence

rate of identified cases ranges from 1 in 400 to 1 in 1000, and the age of diagnosis also varies. The ADPKD hallmarks include progressively enlarging kidney cysts, hypertension, and, ultimately, ESRD, making it the most prevalent monogenic kidney disease worldwide [4].

Historically, the primary focus of treatments for ADPKD was symptom management, with limited impact on cyst development and subsequent kidney edema [5]. However, we can occasionally use specific pharmacological agents, such as tolvaptan, to slow down the growth rate of cysts [6]. In March 2014, tolvaptan received its initial approval for individuals with autosomal dominant polycystic kidney disease (ADPKD) in Japan [7]. Then, on 24 April 2018, the U.S. Food and Drug Administration (FDA) approved tolvaptan, making it the first medication available in the United States for adult patients diagnosed with autosomal dominant polycystic kidney disease (ADPKD) [8].

However, administration of this medication can result in adverse consequences, including an increase in aminotransferase and bilirubin levels, which have the potential to cause liver damage [9]. Only patients with a high-risk rapid progression profile are eligible to receive this medicine due to the potential risk. The total kidney volume (TKV) is the gold standard for determining whether someone has a high-risk profile of rapid progression or not [10,11]. But one big problem with using TKV in clinical practice is that it's challenging to make accurate lines around the kidneys, and it takes a lot of time (45–90 min per patient) for skilled and experienced staff to measure the TKV [12]. Therefore, there is a need for alternative methods to accurately and efficiently classify high-risk ADPKD patients to ensure appropriate treatment and management strategies.

Based on these challenges, we propose a new methodology that utilizes artificial intelligence (AI) to predict the high-risk profile of ADPKD patients using data sources other than TKV. Previous studies that used electronic health records (EHR) combined with AI showed significant results for prediction, diagnosis, and the ability to improve operations and streamline services [13,14]. Numerous studies have applied artificial intelligence (AI) in the medical field. For instance, in 2022, the You Only Look Once (YOLO) version four tiny model was adapted for wireless endoscopic images to detect critical diseases such as cancer [15]. Additionally, these processes are increasingly integrating explainable AI (XAI) techniques. In 2023, Mukhtorov et al. combined the ResNet model (ResNet152) with GradCAM to sort endoscopic images into groups. They performed this by making heat maps that could be used to explain the presence of dyed polyps and other problems [16].

Furthermore, there is specialized research in the domain of kidney disease that has implemented AI for processing electronic health records (EHR). In 2021, Bernardini and colleagues developed a semi-supervised multitask learning (SS-MTL) approach to predict short-term kidney disease outcomes. The SS-MTL method accurately distinguished between early-stage chronic kidney disease (CKD) Stage I (normal cases) and more advanced, higher-risk stages by analyzing temporal changes in EHR data [17]. Prior to this, Njoud et al. (2019) used AI to diagnose CKD in its early stages based on health condition data from 400 individuals, 250 of whom were diagnosed with CKD [18]. Building on the same dataset, Raihan et al. (2023) employed the eXtreme Gradient Boosting (XGBoost) algorithm for CKD prediction, incorporating explainable AI (XAI) techniques through Shapley Additive Explanations (SHAP) to assess feature importance [19]. Despite these advances, there are still challenges in utilizing AI for healthcare data. Certain AI methods require substantial amounts of data, and obtaining clean, reliable, and useful data for healthcare research can be particularly difficult [12].

Aside from that, the problem with AI is the common trade-off between accuracy and interpretability. Knowledge-based AI systems tend to prioritize performance over understandability and explainability, while a highly interpretable machine learning model may compromise the quality of its decisions [20–22]. Because of that, we developed the prediction model using the explainable AI (XAI) concept. XAI allows users and parts of the internal system to be more transparent, providing explanations of their decisions in some level of detail. We can evaluate an explanation based on its interpretability and completeness [23]. Therefore, this paper's proposed solution will not only produce a

model that can predict high-risk ADPKD patients but also create an interpretable model for users, specifically doctors, to make them able to understand the rationale behind the prediction model.

2. Methodology

One of the main objectives of this research is to make AI results consumable, understandable, and usable by the end user, in this case, the doctor. Therefore, we incorporate a user-centered explainable AI approach as outlined by Ribera and Lapedriza. They argued that explanations are multifaceted and cannot be captured by a singular, static account, meaning they must be tailored to the specific user profile [24]. They identified three key stakeholders who need to collaborate to improve the explainability of AI: the developer/AI researcher, domain experts, and lay users/end users [24].

To support this, we adopt the CRISP-DM (CRoss-Industry Standard Process for Data Mining) framework. CRISP-DM is a well-established and widely accepted process model in data mining operations [25]. Additionally, CRISP-DM is well suited to accommodate user-centered explainable AI. The framework serves as a common point of reference for discussing data mining processes and enhances the understanding of key data mining concepts among all stakeholders, particularly end users [26]. The CRISP-DM framework consists of six distinct phases, though the order of these phases is flexible [26]. The six phases are business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

To ensure that user-centered explainable AI is achieved, we map each phase of the CRISP-DM framework to the relevant stakeholders involved in each process. In this research, only five CRISP-DM processes are conducted, as the deployment process is outside the scope. The three main stakeholders are the author of this paper as the AI researcher, nephrologists as the domain experts, and doctors in general as the end users. Table 1 maps the five processes and the activities conducted by each stakeholder.

Table 1. Mapping of processes and stakeholder roles and responsibilities.

Process Stage	AI Researcher (the Author)	Domain Experts (Nephrology Doctor)	End Users (General Doctor)
Business Understanding	Gathers requirements and defines the business problem from the AI perspective.	Provide input on medical requirements, domain knowledge, and expected outcomes.	-
Data Understanding	Analyzes the dataset, identifies data quality issues, and ensures data are suitable for AI tasks.	Offer insights into interpreting the medical data and validating the features.	-
Data Preparation	Cleans, transforms, and preprocesses the data for modeling.	Provide feedback on the clinical relevance of data transformations or feature engineering.	-
Modeling	Builds and trains the AI model. Implements XAI.	Validate the medical accuracy of the model's results and guide any medical assumptions.	-
Evaluation	Evaluates the AI model's performance and accommodates human evaluation.	Perform human evaluation using the proposed explainability matrix.	Perform human evaluation using the proposed explainability matrix.

The next section will divide the paper into three sections, each explaining the implementation of each CRISP-DM phase, with the exception of business understanding and

deployment. The Section 1 provides a brief explanation of the business understanding, while the deployment step is out of scope. We perform evaluation as the last phase without continuing to the deployment phase.

3. Data

3.1. Data Understanding

The Critical Path Institute (C-Path) has organized the raw data for the modeling process from the Polycystic Kidney Disease Outcomes Consortium (PKDOC) database. The data have been standardized and consolidated into a uniform format utilizing the Clinical Data Interchange Standards Consortium (CDISC) Standard Data Tabulation Model (SDTM) framework. The data span roughly 70 years of patient records. The universities included in the list are the University of Colorado-Denver, Mayo Clinic, and Emory University. The database includes information from 2498 individuals, whose ages at the start of the study ranged from 0 to 84 years, with an average age of 35.9 and a middle value of 37. Most of the participants, specifically 81%, are of Caucasian ethnicity, while 59% of them are female [27].

Figure 1 illustrates the process of understanding and preparing data. The first step in this stage is acquiring the data. On 29 November 2023, 44 CVSs of data were downloaded. We imported the data into the RDBMS, specifically PostgreSQL. To understand both the structure and semantics of the data, the Study Data Tabulation Model Implementation Guide (SDTMIG): Human Clinical Trials Version 3.4 (SDTM 3.4) [28] should be a reference. In addition, we conducted a consultation with the C-Path team, the data provider, and a nephrologist, a domain expert, to dive deeply into the data semantics.

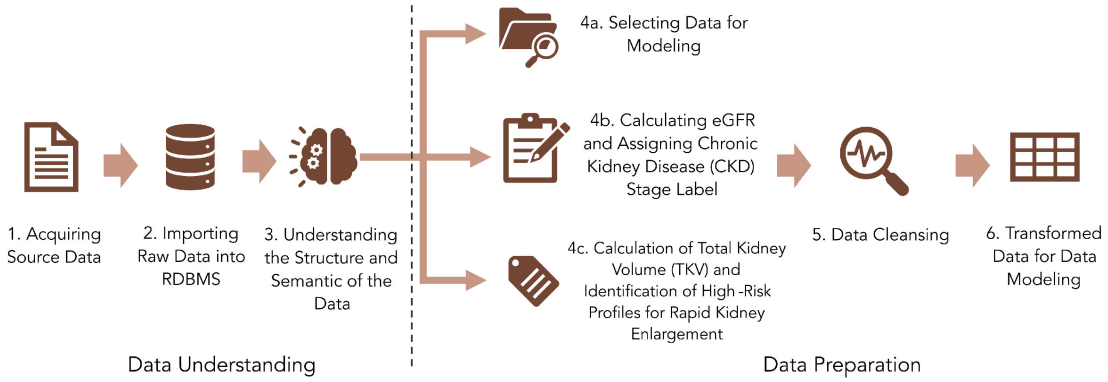


Figure 1. Process for data understanding and preparation.

3.2. Selecting the Data

We selected the data for modeling based on consultation and understanding of the referenced data. Then, we made the selection based on the data availability (not null value) and their potential relationship with the high-risk profile indicator. Figure 2 illustrates the data selection process; 10 table-selected data points are used for the modeling process.

Input	Process	Method	Output
44 CSV	Understanding the structure and semantics of the data.	Read the SDTM 3.4 manual.	10 CSV
		Refer to the PKDOC documentation.	
		Consult with C-Path team.	
	Check the data's availability (the majority of the data is not null).	Query in RDBMS.	
	Check the potential relationship with the high -risk profile indicator.	Consult with a nephrologist	

Figure 2. Data selection process.

Table 2 enumerates all the selected downloaded CSVs, provides a brief description of the data, summarizes the statistical data, and explains the selection criteria for the modeling process.

Table 2. Table of selected data.

File Name/Num. of Rows/Num. of Attributes	Brief Description and Reasons Selection	Statistical Information
DM/2498/29	Demographics of patients as subject observations. Describe personal information as the reference for all tables. Age, sex, race, and ethnicity are used as a reference for subject observation.	All 2498 rows have a unique user id.
PR/764/40	List of the procedures that have been performed on the patients. The procedure is used because this might be related to how severe the patient is.	Only 326 patients (around 13%) have a procedure record. There are 24 intervention names that are categorized into 4 categories: liver, cyst, stone, and blood procedure.
SU/7667/43	Information about the substance used regularly by the patients. The daily substance may be associated with the patient's daily lifestyle, potentially influencing its severity.	65% of the patients have records (1623 of 2498 patients). There are five unique substance categories: alcohol, tobacco, water, decaffeine, and caffeine.
CE/31449/37	A domain for events containing clinically significant occurrences that are not adverse events. The clinical events describe the health conditions of patients.	93% of patients have a CE record (2325 of 2498). There are 46 unique clinical event categorizations. One patient can have multiple clinical events (multiple rows).
HO/1698/29	Record the inpatient or outpatient event, such as a hospitalization or rehabilitation event, for the patient. Although there is not enough information about whether the hospitalization event is because of the ADPKD or not, this information might reflect the severity of the patients.	The inpatient event records only hospitalization. Only 34% of patients have a HO record (326 of 2498)
MH/27464/37	Record the patient's prior medical history as reported. Medical history documents the patient's complete range of problems and all the diagnoses that have been established.	Patient's primary diagnosis is ADPKD. There are 39 unique medical history terms, such as hypertension, migraine, and so on.
VS/308484/35	Record the vital signs of patients, such as temperature, height, and weight. The latest height measurement is the only vstest used; these data are used to calculate hTKV (height-TKV) as a basis to define the high-risk kidney enlargement'classification.	There are eight unique vital signs category recorded: body mass index, body surface area, diastolic blood pressure, heart rate, height, pulse rate, systolic blood pressure, and weight.
FH/4929/29	Specifically contain records of the family history of ESRD (end-stage renal disease) or ADPKD. Since ADPKD is an inherited disease, the affected relative information becomes important.	There are 1808 patients (72.3%) reported to have a family history of ESRD and/or ADPKD. There are 19 distinct subject categories, such as "FATHER" and "MOTHER".
LB/196527/51	Record the patient's laboratory test results. We do not use all of the test results; we only use data to calculate the estimated glomerular filtration rate (*eGFR*). The *eGFR* uses for assessing the presence and degree of renal disease.	There are 3 categories of lab tests (chemistry, hematology, and urinalysis) and 34 unique lab tests or examinations reported (such as creatinine, potassium, and so on).
MP/29678/42	Specifically record the kidney measurements such as volume, width, depth, and mass. The kidney measurement is needed to calculate the total kidney volume (TKV). TKV is an essential calculation to define the patient's high-risk kidney enlargement profile.	95% of patients (2371 of 2498) have a kidney measurement record. The object is measured and categorized as right kidney, left kidney, or bilateral (both kidneys). The method of measurement is either ultrasound, MRI, or CT.

3.3. Calculated Attributes

Other than using the raw data stored in PKDOC database, there are two attributes that are created based on the raw data calculation. The two attributes are the CKD stage and the ground truth for high-risk enlargement kidney profile classification. To populate these two attributes, we calculated the attributes *eGFR* (estimated glomerular filtration rate) and hTKV (height-adjusted total kidney volume).

3.3.1. CKD Stage

The first calculated attribute is the CKD stage classification. Polycystic kidney disease (PKD) is a form of chronic kidney disease (CKD), with the CKD stage label utilizing the estimated glomerular filtration rate (*eGFR*) as a parameter. Research reveals significant negative correlations between the total kidney volume (TKV) and *eGFR*. Lower *eGFR* values are associated with increased TKV growth [29]. By this statement, we can conclude that *eGFR* might be one of the crucial attributes to identify the high-risk profile. Figure 3 illustrates the process to obtain the CKD stage classification attribute.

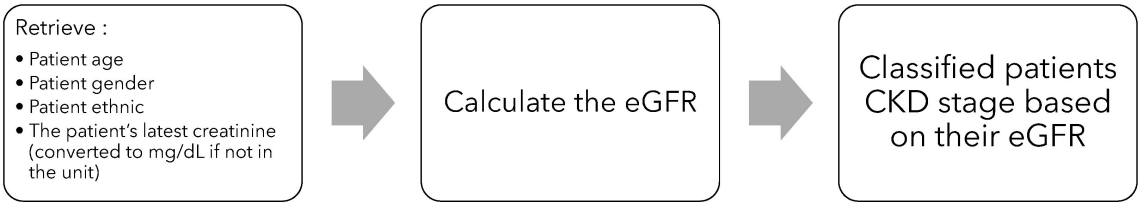


Figure 3. Incorporation of CKD stage classification attribute.

We calculate the *eGFR* using the CKD-EPI Equation (1). The equation is as follows: *Scr* represents serum creatinine in mg/dL, κ is a constant of 0.7 for females and 0.9 for males, and α is a constant of -0.329 for females and -0.411 for males [30]:

$$eGFR = 141 \times min\left(\frac{Scr}{\kappa}, 1\right)^{\alpha} \times max\left(\frac{Scr}{\kappa}, 1\right)^{-1.209} \times 0.993^{Age} \times 1.018[if\ female] \times 1.159[if\ black] \quad (1)$$

We determine the *eGFR* and then attribute it to the five stages of CKD. The *eGFR* test result, which estimates the kidneys' effectiveness in removing waste and excess fluid from the bloodstream, determines the stages. A lower *eGFR* means renal illness worsens and the kidneys' efficiency decreases. An *eGFR* result of 90 or above indicates stage 1, an *eGFR* ranging from 60 to 89 indicates stage 2, and an *eGFR* result of 45 to 59 indicates stage 3a. We classify Stage 3b as occurring when the estimated glomerular filtration rate (*eGFR*) falls between 30 and 44. When the *eGFR* falls between 15 and 29, we assign Stage 4. When the *eGFR* falls below 15, we designate Stage 5 [31].

The data to calculate *eGFR*, including the creatinine result, age, gender, and ethnicity, are already available in the raw data. When calculating creatinine levels for patients with multiple data points, we use only the most recent data point. We determine the creatinine level by analyzing the data in the LB table, specifically the 'Creatinine' lab test under the 'Chemistry' category, which involves examining blood samples. The original unit of measurement for creatinine is "mg/dL". Prior to calculating the *eGFR*, it is necessary to perform a conversion if the data are expressed in units other than mg/dL. Of the total 2498 patients, 85.1%, namely, 2127 individuals, possess sufficient data to compute the estimated glomerular filtration rate (*eGFR*).

3.3.2. High-Risk Profile Classification

The high-risk profile classification is the target class for the modeling process. To obtain this value, we need to calculate and perform a classification process using the raw data. Figure 4 depicts the process of populating the high-risk profile classification.

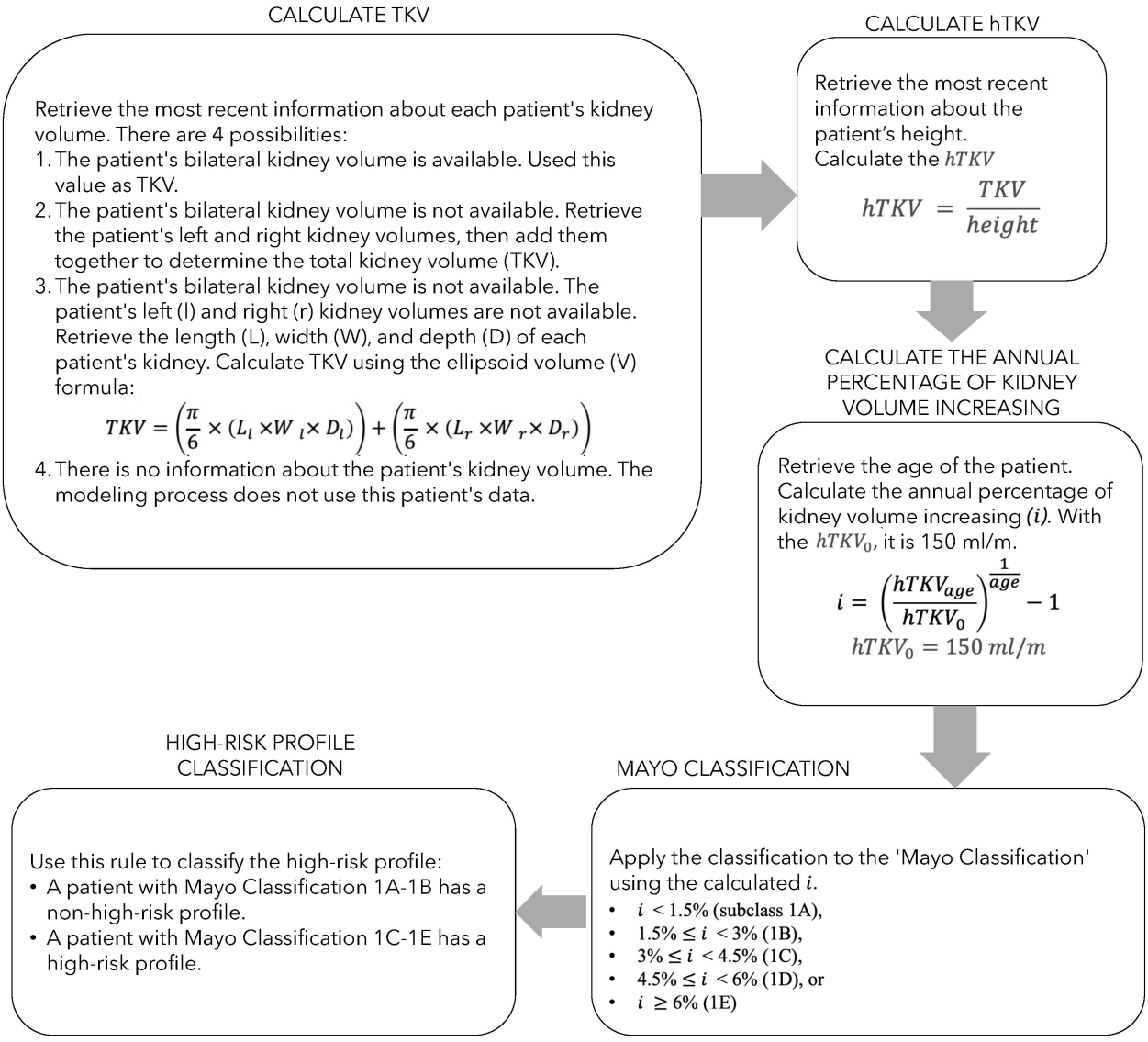


Figure 4. Process of high-risk profile classification.

We use the hTKV calculation to classify patients with ADPKD as having a high-risk enlargement of their kidney profile. We group patients into sub classes 1A–1E based on their hTKV measurement. The TKV, or the kidney total volume from both sides, measures the hTKV. To adjust for the patient's height, we divide the TKV by the patient's height. We find hTKV ranges for each age group, starting at 150 mL/m and dividing the annual percentage increases into five groups: less than 1.5% (subclass 1A), 1.5% to 3% (1B), 3% to 4.5% (1C), 4.5% to 6% (1D), or more than 6% (1E) [32]. This classification, known as the Mayo Classification, 1C–1E, is for patients with a high risk of rapid progression of ADPKD [33]. This rule assigns a binary attribute named "high risk profile", classifying patients with a Mayo classification of 1A or 1B as '0' (non-high-risk profile), and those with a Mayo classification of 1C, 1D, or 1E as '1' (high risk profile). We conduct Mayo classification research on patients who are over the age of 15. As a result, the high-risk profile attribute only populates patients who are at least 15 years old.

We calculate the annual percentage of kidney volume using patient volume data from the MP table. The data, obtained from multiple sources and spanning over a decade, may vary in completeness for each patient. To obtain the TKV patient data, we implement several regulations as shown in Figure 4 in the 'Calculate TKV box. You must express the volume data in milliliters (mL) and the height in meters (m). If you use any other units,

please convert them accordingly. Out of 2498 patients, 89.7%, or 2240, had sufficient data to be classified as either having a high-risk or non-high-risk profile.

3.4. Cleansing the Data

After selecting the data and populating new attributes for modeling, we need to perform the cleansing step. Certain cleansing processes are necessary to ensure the quality and reliability of the data used for analysis.

1. High Variance Value: We encounter an issue where multiple variations convey the same meaning due to data sourced from multiple channels over the years. This issue can be caused by typographical errors or variations in term usage. Some typographical errors exist in the raw data, such as 'Caffien', which should have the same meaning as 'Caffein'. There are also differences in using the term, such as when some patients have medical records for 'Tuberculosis' and others have 'TB', which actually have the same meaning. We anticipate that this high variance value will have an impact on the modeling accuracy, so we manually set the grouping value for some attributes to minimize the variance. For instance, we classify substances called 'Caffien' and 'Caffein' using the same feature as 'Substance_Caffein'. The doctor, with his domain knowledge, assists in this manual grouping process to prevent misinterpretation.
2. Handle Null Data: Null data are a critical aspect to avoid in the modeling process since their existence can have a substantial impact on the modeling outcomes. Therefore, it is crucial to handle this issue. In this research, there are two approaches to handling it: (1) populate it with an exception categorical like 'UNKNOWN', or (2) do not use the patient data in the modeling process. When the data are categorical, not numeric, and have a prior exception category, we choose the first approach. Example: Out of 2498 patients, 205 (8.21%) have a null value in their 'race' feature. Since there are already 7 categories of race, including 'UNKNOWN', we update the 205 patients with the 'Null' race to 'UNKNOWN'. Unlike approach 1, we use approach 2 when the feature is numeric in nature, allowing us to not estimate the value. There are 371 patients who have no 'Creatinine' data to calculate their CKD stage; unfortunately, we are not included in the modeling process to reduce the bias.
3. Irregular data: The raw data are derived from patient registries, not from clinical studies. The timing of the visits is irregular, and not all tests or measurements are conducted consistently [27]. This causes doubts about the reliability of time-related data points. Figure 5 depicts a snapshot of the medical history of one patient. Both the first and second rows record the same medical history, 'Abdomen Protube', at the same visit time, 'visit 1', and on the same study day for history collection (mhdy feature), '1'. Therefore, we can assume these data are duplicates. Hence, instead of relying on a count-based approach, we utilize the max aggregate method. For instance, we set the 'Hypertension' column to 1 if a patient has at least one recorded instance of hypertension. On the other hand, if there are no such records, we set the column to 0.

The same visited # with the same medical record history

studyid	DOMAIN	usubjid	mhseq	mhterm	visit	mhdy	mhcat
PKD1	MH		01 4	Abdomen Protube	VISIT 1	1	GENERAL
PKD1	MH		01 9	Abdomen Protube	VISIT 1	1	GENERAL
PKD1	MH		01 1	ADPKD	VISIT 1	1	PRIMARY DIAGNOSIS
PKD1	MH		01 5	ADPKD	VISIT 1	1	PRIMARY DIAGNOSIS
PKD1	MH		01 7	Anorexia	VISIT 1	1	GENERAL
PKD1	MH		01 3	Flank Pain	VISIT 1	1	GENERAL
PKD1	MH		01 2	Hypertension	VISIT 1	1	GENERAL
PKD1	MH		01 6	Hypertension	VISIT 1	1	GENERAL

Figure 5. Snapshot of patient medical history.

3.5. Transformed Data

To condense the intricate dataset into a more streamlined format, we transform all the relevant data into a single tabular structure. Appendix A lists the tabular data used for the modeling process. The feature that is not in the binary type is encoded in the modeling process. There are 100 feature variables and 1 target variable. In total, 1779 of 2498 patients' data are used in the modeling process; the other 494 are not used because there are insufficient data to calculate the CKD stage or/and TKV data to be classified as high-risk profiles or not. We label 1254 patients as high-risk profiles and 525 as non-high-risk profiles.

3.6. Dataset Diversity Limitation

Our study acknowledges several limitations related to the demographic of the dataset. Out of 1779 patients, we analyze the available demographic data to identify potential biases that may affect the model's performance. As Franklin et al. (2024) noted, sociodemographic characteristics—including race, ethnicity, gender, age, and insurance status—can lead to unequal treatment based on specific demographics, potentially influencing the outcomes of predictive models [34].

Figure 6 and Table 3 display the distribution of four key demographic features: sex, age, race, and ethnicity. From this distribution, it becomes clear that there is significant inequality across almost all features. For example, the age data follow a normal curve, but only patients over the age of 15 are included in accordance with the Mayo Classification. In terms of sex, the dataset is imbalanced, with more than half of the patients being female.

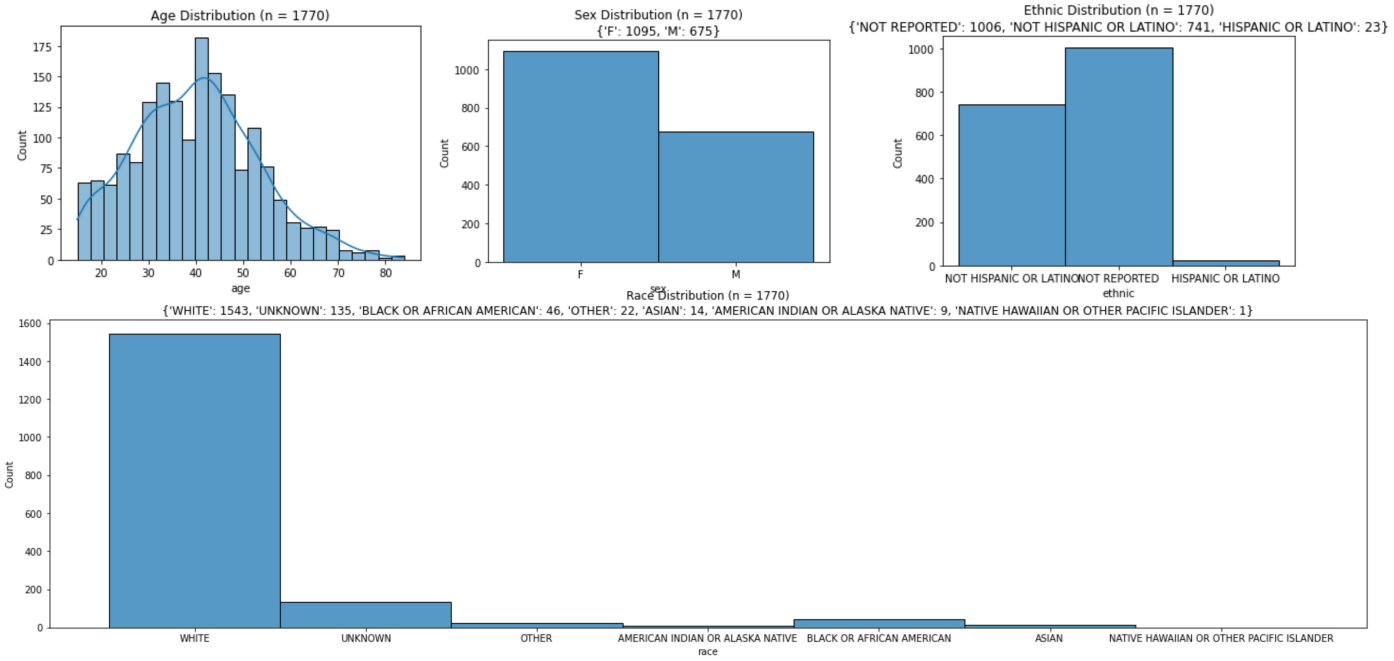


Figure 6. Demographic data distribution chart.

The most pronounced imbalance, however, occurs in the race and ethnicity data. Ethnic information is largely unreported, and a striking 87.18% of the patients belong to the white race. This over-representation of one demographic group raises concerns about the potential bias in our model, as it may lead to predictions that are more accurate for one group while under-performing for others. Although excluding demographic data might seem like an option to avoid bias, this would not be advisable, as research indicates that these characteristics are strongly correlated with the disease outcomes being studied.

Therefore, while our research presents valuable findings, we recognize that the lack of diversity in the dataset poses a limitation. Future research should aim to include more

balanced and representative datasets to ensure that AI-driven models in healthcare are equitable across diverse populations.

Table 3. Demographic data distribution.

Variables	n	%
Age		
15–20	128	7.23%
21–30	312	17.63%
31–40	473	26.72%
41–50	489	27.63%
51–60	250	14.12%
61–70	91	5.14%
71–80	22	1.24%
81–85	5	0.28%
Sex		
Female	1095	61.86%
Male	675	38.14%
Ethnic		
Not Hispanic or Latino	741	41.86%
Hispanic or Latino	23	1.30%
Not Reported	1006	56.84%
Race		
White	1543	87.18%
Black or African American	46	2.60%
Asian	14	0.79%
American Indian or Alaska Native	9	0.51%
Native Hawaiian	9	0.51%
Other	22	1.24%
Unknown	135	7.63%

4. Modeling

The experimental scenario illustrated in Figure 7 comprises several key components, which are detailed below:

1. Data Splitting: The transformed data are split into training and testing sets to ensure robust model evaluation. The data are divided in an 80:20 ratio, where 80% of the data is used for training the model and 20% is reserved for testing during the evaluation phase.
2. Dataset Definition: The experimental modeling utilizes three distinct datasets, the details of which are provided in the following section. These datasets are curated to reflect diverse features relevant to high-risk patient identification.
3. Handling Imbalanced Data: As discussed earlier, the ratio of high-risk to non-high-risk patient profiles in the dataset is 2.4:1. This class imbalance may cause the model to be biased toward the majority class, potentially leading to over-fitting, where the model performs well on the training data but poorly on unseen test data. To mitigate this issue, two techniques were employed.
4. Model Training: Seven machine learning algorithms are selected for this study. The rationale for selecting these algorithms, along with the experimental results, is discussed in detail in a later section.
5. Model Performance Metrics: The models are evaluated using several performance metrics, including accuracy, precision, recall, F1 score, and the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). AUC-ROC is chosen as the primary performance metric, as it provides a robust measure of performance for both imbalanced and balanced datasets, being less influenced by the distribution of classes [35].
6. Hyperparameter Tuning: The model with the highest AUC score is further optimized through hyperparameter tuning. Optuna, a hyperparameter optimization framework,

is utilized to systematically search for the best set of parameters, improving the model's performance.

7. Explainable AI (XAI) Approach: To enhance the interpretability of the selected model, two local model-agnostic XAI techniques are applied: Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME). These methods are chosen because they can be applied to any machine learning model, regardless of its architecture.
8. Explainability Evaluation: The primary objective of this research is to improve experts' interpretability of the AI model's predictions. Therefore, a human evaluation process is conducted. A custom evaluation matrix, focusing on the explainability of the model's predictions, is developed and distributed to medical experts (doctors) for assessment.

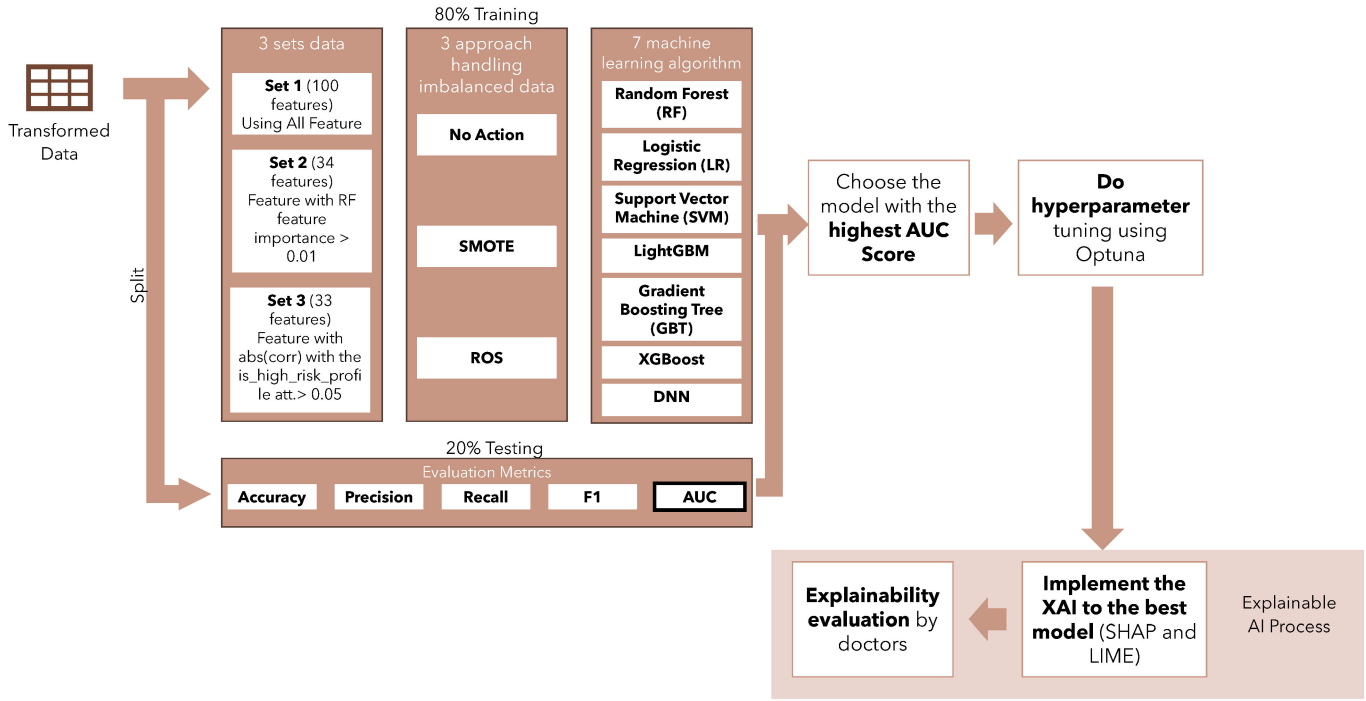


Figure 7. Experimental scenario.

The following sections elaborate on the details of each component of this process.

4.1. Define Datasets

There are three sets of data prepared for the training process. The first set uses all 100 features as mentioned in Appendix A. The second set uses the Random Forest feature importance method. We only use the feature importance score bigger than 0.01 as depicted in Figure 8. The third set uses only features that have an absolute correlation score greater than 0.05 between the feature and the target variable (is_high_risk_profile). Table 4 lists the features used in the second and third sets; there are 17 common features between the two sets (bold words in the table).

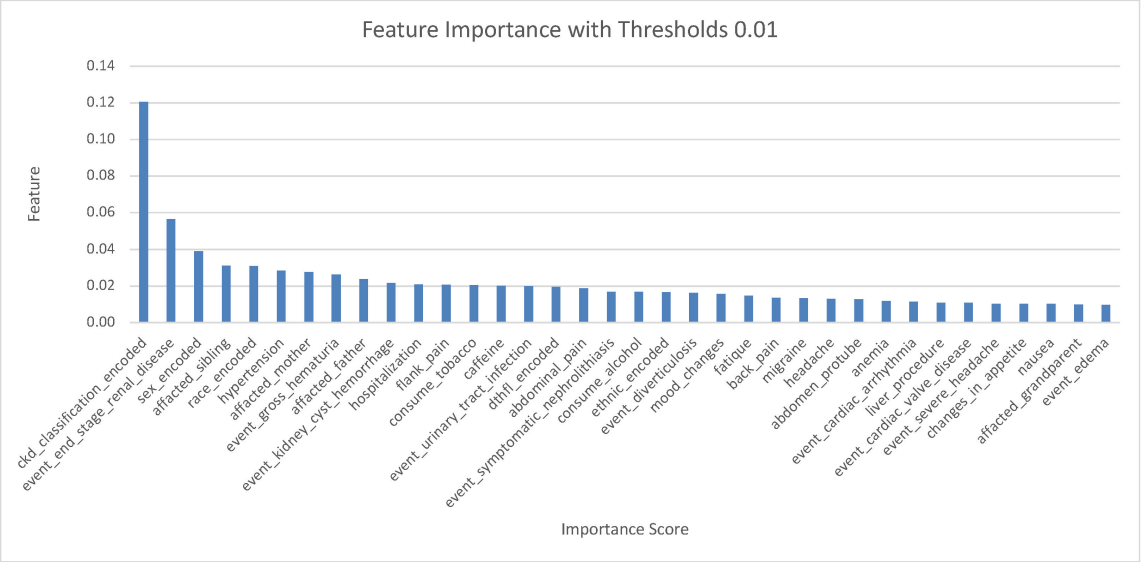


Figure 8. Feature importance.

Table 4. Set for modeling process.

Set	Feature [1]
Set 2 (34 attributes)	**'abdominal_pain'**, **'abdomen_protube'**, 'affacted_father', 'affacted_mother', 'affacted_sibling', **'anemia'**, **'back_pain'**, 'caffeine', **'ckd_classification_encoded'**, 'changes_in_appetite', 'consume_alcohol', 'consume_tobacco', 'dthfl_encoded', 'ethnic_encoded', **'event_cardiac_arrhythmia'**, 'event_cardiac_valve_disease', 'event_diverticulosis', **'event_end_stage_renal_disease'**, **'event_gross_hematuria'**, 'event_kidney_cyst_hemorrhage', 'event_severe_headache', 'event_symptomatic_nephrolithiasis', 'event_urinary_tract_infection', 'fatique', 'flank_pain', 'headache', **'hospitalization'**, **'hypertension'**, **'liver_procedure'**, **'migraine'**, 'mood_changes', **'nausea'**, 'race_encoded', 'sex_encoded'
Set 3 (33 attributes)	**'abdominal_pain'**, **'abdomen_protube'**, **'affacted_father'**, **'anemia'**, 'anorexia', **'back_pain'**, 'blood_procedure', **'ckd_classification_encoded'**, **'consume_tobacco'**, 'diarrhea', 'ethnic_encoded', 'event_aneurysm', **'event_cardiac_arrhythmia'**, 'event_edema', **'event_end_stage_renal_disease'**, 'event_exertional_chest_pain', **'event_gross_hematuria'**, 'event_hepatic_venous_outflow_obstruction', 'event_inguinal_hernia', 'event_intracranial_aneurysm', 'event_nocturia', 'event_non_exertional_chest_pain', 'event_shortness_of_breath_at_rest', 'event_umbilical_hernia', **'flank_pain'**, **'hospitalization'**, **'hypertension'**, 'insomnia', **'liver_procedure'**, **'migraine'**, **'nausea'**, **'sex_encoded'**, 'stone_procedure'

[1] The bold words represent the common features between both sets.

4.2. Handling Imbalanced Data

Imbalanced data, where one class significantly outnumbers another, can adversely affect the performance of machine learning models. In such cases, models often become biased toward the majority class, resulting in poor predictive accuracy for the minority class. This issue is evident in our dataset, where, out of 1779 patients, 1254 are labeled as high-risk profiles, while only 525 are non-high-risk profiles. This imbalance can lead to skewed model predictions, favoring the high-risk group while overlooking the non-high-risk profiles, which may compromise the model's overall effectiveness [36].

To mitigate this, we implement two widely used techniques: SMOTE (Synthetic Minority Over-sampling Technique) and ROS (Random Oversampling) [37]. These algorithms

address the class imbalance by modifying the dataset to achieve a more balanced class distribution. ROS works by randomly duplicating instances from the minority class, thereby introducing additional samples to reduce imbalance. However, while effective, ROS may increase the risk of over-fitting by repeating exact data points from the minority class. On the other hand, SMOTE generates fictional data by considering the similarities in the feature space between existing minority cases. SMOTE not only balances the data but also enriches the dataset with meaningful patterns that may improve model performance [37].

4.3. Model Training

In selecting the machine learning algorithms for model training, two key characteristics are considered. First, since the dataset consists of tabular data, we select six machine learning algorithms known for their effectiveness in handling tabular datasets. These algorithms are widely recognized for their robust performance across various applications involving structured data. The algorithms used include Random Forest (RF) [38], Logistic Regression (LR) [39], Support Vector Machines (SVMs) [40], LightGBM [41], Gradient Boosting Tree (GBT) [42], and XGBoost [43]. Each of these algorithms has been empirically demonstrated to provide competitive accuracy for tabular datasets, particularly in the healthcare domain.

Second, the complexity of the dataset, which includes over 10 features, necessitates the exploration of deep learning approaches in addition to traditional machine learning methods [44]. Deep Neural Networks (DNNs) are chosen for this purpose due to their capability to model complex interactions between features and their strong performance in handling high-dimensional data [45]. DNNs have shown superior ability to capture non-linear relationships and intricate patterns in data with many features, making them well suited for datasets with a large number of variables [46].

We combine three datasets, apply three different approaches for handling imbalanced data, and utilize seven machine learning algorithms. In total, this leads to the execution of 63 distinct experiments, as we test each machine learning algorithm across the different dataset and data balancing combinations. The hyperparameters used in the training model are listed in Appendix B.

4.4. Model Performance Metrics

When evaluating the performance of machine learning models, numerous metrics are available to choose from. Libraries like Scikit-learn provide several commonly used metrics, including accuracy, precision, recall, and F1 score. Accuracy, which is calculated as the ratio of correct predictions to the total number of predictions, is not an ideal metric for medical applications, where data are often highly imbalanced [47].

Sensitivity (recall) measures the proportion of actual positive cases that are correctly identified, while precision measures the proportion of predicted positive cases that are truly positive. Although these metrics are used in some research, they have limitations. Specifically, precision and recall may not fully capture model performance due to the large background class in imbalanced datasets, which can inflate specificity [48].

The F1 score, which combines both precision and recall, provides a more balanced measure, particularly in scenarios where the positive class is crucial. It assesses the overlap between the predicted results and the actual ground truth [48]. Similarly, the Receiver Operating Characteristic (ROC) curve graphically illustrates a model's diagnostic ability across different thresholds, and its Area Under the Curve (AUC) quantifies overall performance [49].

In the medical domain, metrics like the F1 score and ROC AUC are often more informative than accuracy alone [48]. AUC, in particular, is a robust metric that performs well in both balanced and imbalanced datasets because it is less affected by the distribution of class instances [35].

Table 5 lists the ten models with the highest AUC scores. Performance metrics for all 63 models can be found in Appendix C.

Table 5. Ten models with the highest AUC score.

Set Attributes	Training Algorithm	Imbalanced Method	Accuracy	Precision	Recall	F1	AUC
Set 3 [1]	**XGBoost**	**SMOTE**	**0.715**	**0.682**	**0.702**	**0.687**	**0.702**
Set 3	SVM	SMOTE	0.681	0.672	0.699	0.666	0.699
Set 3	LightGBM	SMOTE	0.703	0.672	0.691	0.676	0.691
Set 3	SVM	ROS	0.652	0.662	0.688	0.643	0.688
Set 2	GBT	ROS	0.678	0.661	0.687	0.660	0.687
Set 3	GBT	ROS	0.670	0.660	0.686	0.654	0.686
Set 3	LR	ROS	0.675	0.661	0.685	0.657	0.685
Set 3	LightGBM	ROS	0.701	0.667	0.684	0.671	0.684
Set 3	DNN	ROS	0.732	0.717	0.732	0.719	0.689
Set 3	RF	SMOTE	0.706	0.668	0.681	0.672	0.681

[1] The bold models are the models with the highest AUC.

From the table, we observe that tree-based machine learning algorithms, specifically XGBoost, achieve higher AUC scores compared to deep learning models. This aligns with the findings of Léo Grinsztajn's 2022 research, which suggests that tree-based algorithms often outperform deep learning models for tabular data. This is due to the characteristics of tabular data, which often contain irregular patterns in the target function, uninformative features, and non-rotationally invariant data, where linear combinations of features fail to accurately capture the underlying information [50].

Additionally, the performance metrics indicate that the feature selection process effectively improves model accuracy. Datasets with more features tend to include irrelevant and noisy data, which can negatively impact model performance. Notably, the dataset with the highest number of features, Set 1 with 100 features, does not appear among the ten models with the highest AUC scores, further emphasizing that larger feature sets do not necessarily lead to better performance.

4.5. Hyperparameter Tuning

Based on the previous section, the model with the highest AUC score is the one trained on Set 3, which uses SMOTE to handle imbalanced data and is trained using the XGBoost algorithm. This model is then further optimized using the Optuna framework. Optuna is widely recognized for its efficiency in automating hyperparameter optimization, particularly through its Tree-structured Parzen Estimator (TPE) and other advanced algorithms, which allow for faster and more accurate tuning compared to traditional methods like grid or random search [51].

We optimize the XGBoost model using seven parameters: 'lambda', 'alpha', 'max_depth', 'eta', 'gamma', 'colsample_bytree', and 'min_child_weight'. After tuning, the AUC score is increased by 0.1 as shown in Table 6.

Table 6. Model performance after tuning.

Hyperparameter Tuning	Accuracy	Precision	Recall	F1	AUC
'lambda': 3.9957070357300086, 'alpha': 2.799141734671868, 'max_depth': 7, 'eta': 0.08770989189740201, 'gamma': 0.1671936607245508, 'colsample_bytree': 0.9400135473143985, 'min_child_weight': 9	0.709	0.686	0.712	0.689	0.712

As the primary metric in this research is the Area Under the Curve (AUC), the hyperparameter tuning process is centered around maximizing the AUC score. After tuning, the AUC score, along with other key performance metrics such as precision, recall, and F1 score, show improvement. However, it is observed that the accuracy of the model slightly decreases. This outcome is not uncommon in machine learning, especially when

dealing with imbalanced datasets. Accuracy, while often used as a general metric, can be misleading in such cases because it does not account for the distribution of classes. A high accuracy can be achieved even if the model is performing poorly on the minority class, which is why it is less informative than metrics like AUC, precision, and recall in these situations. The gains in AUC, precision, recall, and F1 score suggest that the model is now more capable of making balanced predictions, which is a more desirable outcome in many real-world applications.

4.6. Explainable AI

Many papers have previously described the concept of AI using various terms, yet they all concur that the primary goal of XAI is to enable users to understand the rationale behind a model's decision-making process. From multiple survey papers of the implementation of XAI, there are several challenges that present opportunities for further research and potential solutions. The problem at hand is the common trade-off between accuracy and interpretability in AI systems [20–22].

Then, in 2016, Ribeiro, Singh, and Guestrin introduced the concept of separating explanations from the model itself, known as model-agnostic interpretability. This approach provides a flexible and generic framework for interpretability, independent of the underlying machine learning model [52]. Two of the most prominent model-agnostic techniques are Local Interpretable Model-Agnostic Explanations (LIME) [53] and SHAP (Shapley Additive Explanations) [54].

The LIME algorithm generates a local surrogate model by training models that approximate the predictions of the underlying black-box model for specific instances. LIME focuses on local surrogate models to provide explanations for individual predictions [55]. This flexibility makes LIME highly adaptable to various machine learning algorithms, and it is particularly effective for processing tabular data.

On the other hand, the SHAP framework defines a class of additive feature importance methods and demonstrates the existence of a unique solution within this class that satisfies certain desirable properties [54]. SHAP is widely regarded as a powerful tool for Explainable AI (XAI) due to its mathematical rigor, consistency, and ability to provide interpretable explanations for any machine learning model. Furthermore, it maintains flexibility across different types of models, making it a highly versatile approach to interpretability.

Both LIME and SHAP provide local explanations, meaning they are capable of interpreting the conditions of individual data points. In this study, two cases are selected: one with a high probability result, and one with a near 50:50 probability split. Visualizations for the extreme cases are presented in Appendix D.

The figures below (Figures 9–11) present the prediction results for a borderline case involving a 33-year-old female patient diagnosed with polycystic kidney disease. She has a family history of kidney disease, as her father is also affected. The patient experiences a variety of symptoms, including gross hematuria, hypertension, shortness of breath at rest, edema, nausea, inguinal and umbilical hernias, flank pain, intracranial aneurysm, anorexia, chest pain, back pain, and abdominal pain. However, she does not exhibit symptoms such as migraine, hepatic vein obstruction, diarrhea, insomnia, cardiac arrhythmia, anemia, or other aneurysms. The patient has never been hospitalized or undergone procedures related to kidney stones, cysts, or liver issues. Based on her *eGFR*, she is classified as being at stage 2 of chronic kidney disease. Additionally, the patient is an active smoker.

In Figure 9, the model's prediction for this patient is shown without the XAI approach. All patient conditions are numerically encoded for computational processing. A prediction value of 1 indicates that the patient is predicted to be at high risk, while a value of 0 indicates a low-risk prediction.

Subsequently, Figure 10 presents the LIME visualization for this same case, while Figure 11 illustrates the SHAP visualization. Both visualizations effectively demonstrate the contribution of key features to the prediction outcome, highlighting the most important factors influencing the model's decision.

ethnic_encoded	1
event_cardiac_arrhythmia	0
liver_procedure	0
migraine	0
event_hepatic_venous_outflow_obstruction	0
diarrhea	0
insomnia	0
event_end_stage_renal_disease	0
ckd_classification_encoded	1
event_gross_hematuria	1
sex_encoded	0
hypertension	1
event_shortness_of_breath_at_rest	1
event_edema	1
nausea	1
event_inguinal_hernia	1
event_nocturia	1
event_umbilical_hernia	1
affacted_father	1
flank_pain	1
event_intracranial_aneurysm	1
anorexia	1
event_exertional_chest_pain	1
consume_tobacco	1
event_non_exertional_chest_pain	1
back_pain	1
anemia	0
stone_procedure	0
blood_procedure	0
hospitalization	0
abdominal_pain	1
event_aneurysm	0
abdomen_protube	1
Predicted High Risk Profile	1

Figure 9. AI prediction without XAI.

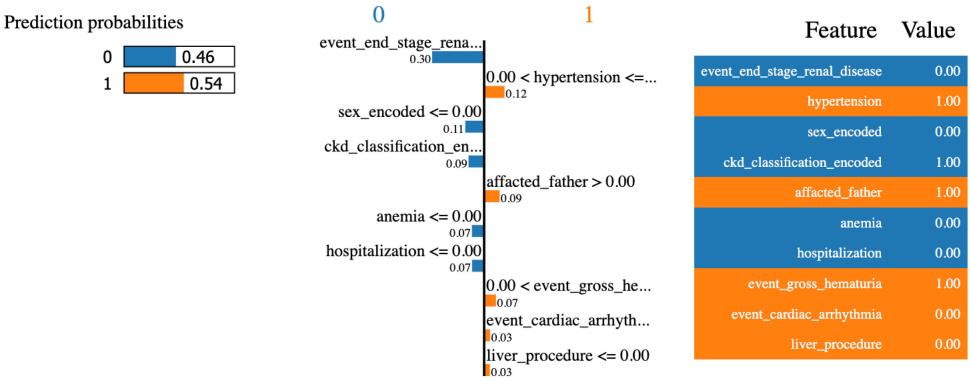


Figure 10. Prediction with LIME visualization.

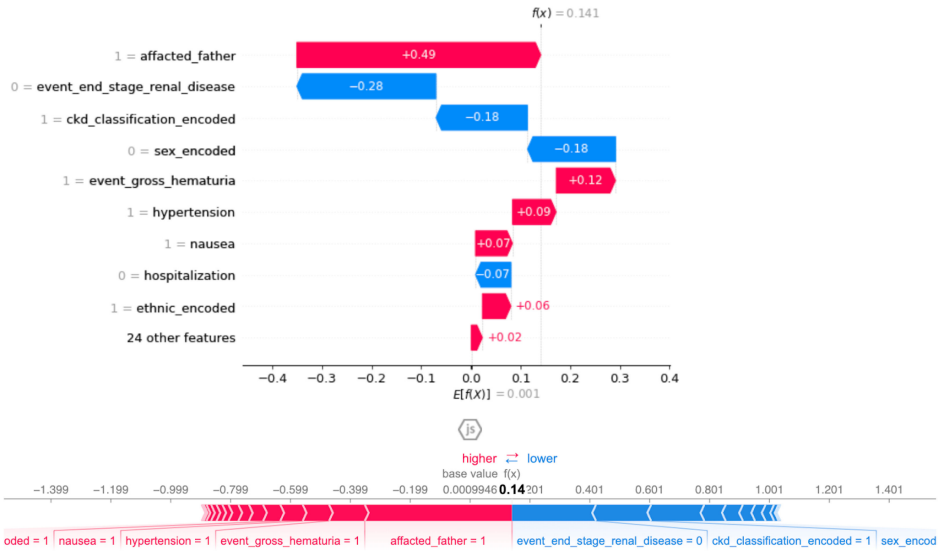
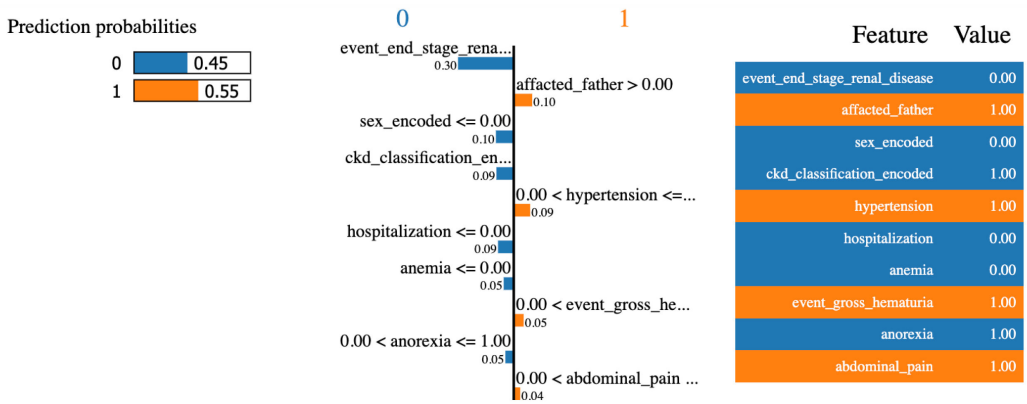


Figure 11. Prediction with SHAP visualization.

To further enhance the explainability of the model, we hypothesize that adding a text component to accompany the graphical visualizations produced by LIME and SHAP would provide a more comprehensive interpretation of the model's predictions. While the visual output from LIME and SHAP effectively highlights the contribution of individual features to the model's decisions, integrating descriptive explanations could offer deeper insights, particularly for stakeholders with less technical expertise.

To achieve this, we propose the integration of a generative language model, specifically ChatGPT, to automatically generate explanatory text based on the results of the LIME and SHAP visualizations. This approach allows for the translation of complex visual data into human-readable explanations that describe how specific features influence the prediction.

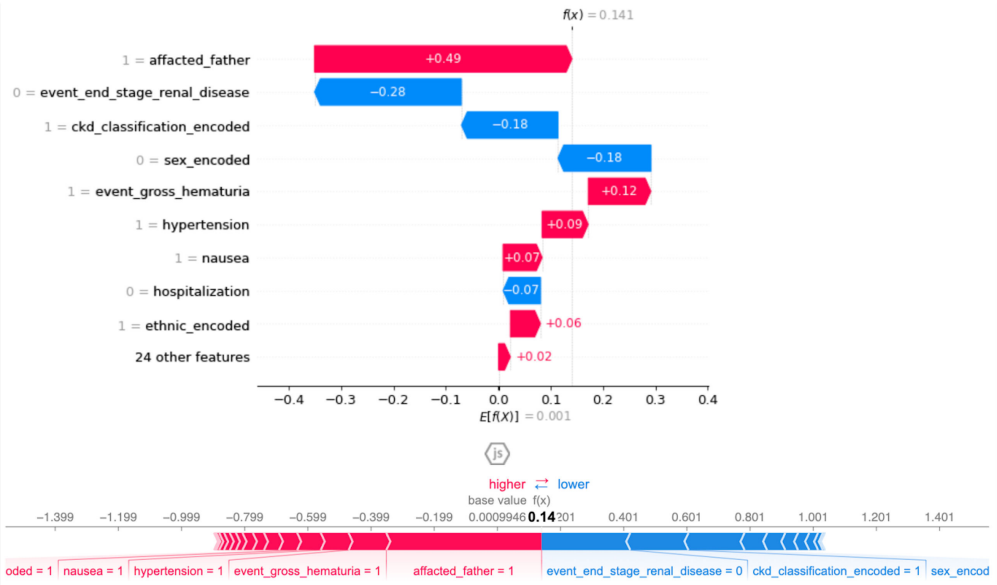
Figures 12 and 13 present the combined visualizations generated using the LIME and SHAP approaches, respectively, alongside explanatory text generated by ChatGPT. For both models, we utilize the same prompt: "Generate an explanation of the prediction result based on the LIME/SHAP output for this patient case", where the patient case details are as outlined in the previous paragraph. The LIME and SHAP visualizations are provided as inputs to the prompt to guide the generation of a coherent and relevant explanation.




```
LIME Analysis:

The LIME results show that the prediction probabilities from the model for the two classes are:

Class 0 prediction (low probability): 46%
Class 1 prediction (high probability): 54%
Features influencing the prediction, along with their importance:

Key factors for Class 1: Hypertension, a family history of kidney disease, gross hematuria,
and several other related health conditions, such as heart-related procedures and arrhythmias,
influence the model's decision to predict a higher probability of a serious condition or
greater severity (Class 1).

Key factors for Class 0: End-stage kidney disease (0.00), gender, and anemia also affect
the prediction for a lower probability (Class 0), though their impact is smaller compared
to factors driving the prediction towards Class 1.

Conclusion: Based on this analysis, while factors such as hypertension and family history
of kidney disease contribute to a higher prediction for a more severe risk class (Class 1),
there are also factors that reduce the risk prediction (Class 0). Overall,
the probability of the patient being classified into the high-risk class is 54%.
```

Figure 12. Prediction with LIME visualization and ChatGPT generated text.

By incorporating the ChatGPT generated text with these visualizations, the figures not only illustrate the importance of specific features but also offer a detailed narrative explaining how these features influence the model's prediction for this particular patient case.




```
Key Factors Increasing Risk (Positive):

Family History of Kidney Disease (affected_father = 1): This is the most significant factor increasing
the risk prediction, with a SHAP value of +0.49.
Gross Hematuria (event_gross_hematuria = 1): This condition has a notable positive impact, with a SHAP
value of +0.12.
Hypertension (hypertension = 1): This factor also contributes positively, with a SHAP value of +0.09.
Nausea (nausea = 1): This condition adds to the risk prediction with a SHAP value of +0.07.
Non-Hispanic Ethnicity (ethnic_encoded = 1): This factor has a small but positive influence, with
a SHAP value of +0.02.

Key Factors Decreasing Risk (Negative):

End-Stage Kidney Disease (event_end_stage_renal_disease = 0): This condition has the most significant
impact in reducing the risk prediction, with a SHAP value of -0.28.
Chronic Kidney Disease Classification (ckd_classification_encoded = 1): This has a considerable
impact in lowering the risk, with a SHAP value of -0.18.
Female Gender (sex_encoded = 0): This factor also reduces risk, with a SHAP value of -0.18.
No Hospitalization (hospitalization = 0): This factor slightly reduces the risk prediction, with
a SHAP value of -0.07.

Conclusion: Based on the SHAP analysis, the strongest risk factors increasing the likelihood of
a serious or more severe condition are a family history of kidney disease, gross hematuria,
and hypertension. However, factors such as the absence of end-stage kidney disease, chronic
kidney disease classification, and female gender contribute to lowering the risk prediction.

In other words, while some factors exacerbate the condition, the model indicates that there are
still significant risk-reducing factors. These results provide a moderate probability of a serious
or more severe condition
```

Figure 13. Prediction with SHAP visualization and ChatGPT generated text.

5. Human Evaluation

As outlined in Section 2 and Table 1, the evaluation process in this research extends beyond standard AI model performance metrics such as accuracy, recall, precision, F1, and AUC. In addition to these mathematical metrics, this study incorporates human evaluation to assess the model's explainability.

A primary goal of this research is to ensure that the AI-generated results are interpretable and usable by end users, specifically doctors in this case. As noted by Doshi-Velez et al. in their work, for machine learning models to be both specific and interpretable, it is essential to conduct human-grounded evaluations [56]. Therefore, the evaluation framework includes feedback from doctors, who assess the level of explainability of the proposed model. To achieve this, a four-step process for human evaluation is implemented, with each step outlined in detail in the following subsections.

5.1. Define Metrics for Explainability

There is no universally agreed-upon definition of interpretability in machine learning, nor is there a clear consensus on how to measure it. To address this, we propose a set of metrics for evaluating explainability. These metrics are derived from the literature on the

goals of Explainable AI (XAI) as discussed by Barredo Arrieta et al. in 2020 [20]. Their work categorizes the goals of XAI based on the intended audience, and for this research, the focus is on domain experts.

The five primary goals of XAI for domain experts are trustworthiness, causality, transferability, informativeness, and confidence. To quantify these aspects of explainability, we adopt a Likert scale, which is effective for measuring subjective perceptions and attitudes [57], such as levels of 'confidence' or 'trustworthiness'.

For each AI model evaluated, five key questions are posed, with responses measured on a Likert scale from 1 (strongly disagree) to 5 (strongly agree):

1. Trustworthiness: I trust the AI model's predictions.
2. Causality: The model's explanations help me understand the cause-and-effect relationships behind the predictions.
3. Transferability: I believe the model's insights can be applied to other patients with similar conditions.
4. Informativeness: The visualizations and explanations provided by the AI model are informative and clear.
5. Confidence: I feel confident in using the AI model's predictions for decision-making.

This structured approach allows us to assess the explainability of the AI models from a domain expert's perspective, ensuring that the models meet critical interpretability goals.

5.2. Selection of Models for Assessment

In this assessment process, multiple models are presented to the users (doctors) for evaluation. The objective is to compare the levels of explainability both before and after the implementation of Explainable AI (XAI). This comparison helps to evaluate how effectively XAI enhances users' understanding of the model's decisions.

Five visualizations are included for assessment by the respondents. The first visualization represents the baseline model, depicted in Figure 9. The subsequent visualizations utilize the LIME approach, with two variations: one with generated text (Figure 12) and one without (Figure 10). The final two visualizations employ the SHAP approach, again with one version including generated text (Figure 13) and another without (Figure 11).

By comparing these five visualizations, we aim to identify the components that contribute to an increased level of explainability for the users. Insights gained from this analysis can inform future enhancements to the models and their interpretability features.

5.3. Conducting the Survey with Users

A significant challenge in conducting surveys involving human participants in AI assessments is the varying levels of understanding regarding artificial intelligence. To address this issue, we provided a video explanation and conducted face-to-face interviews prior to the assessment to ensure that all participants grasped the research objectives clearly.

The survey included an assessment of the five models discussed in Section 5.2. Participants also asked to indicate their preferred visualizations, provide reasons for their choices, and respond to open-ended questions regarding which components they believe need enhancement.

The survey was specifically distributed to doctors. A total of 27 doctors participated in the survey and confirmed their consent for participation. Among the participants, 15 (55.6%) were male, while the remaining respondents were female as illustrated in Figure 14. The average age of the respondents was 31.7 years, with the age distribution depicted in Figure 15.

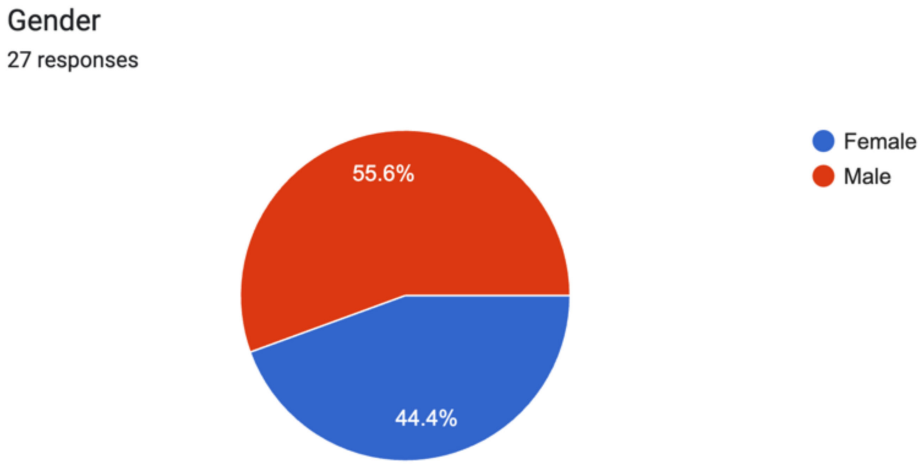


Figure 14. Gender demographics of respondents.

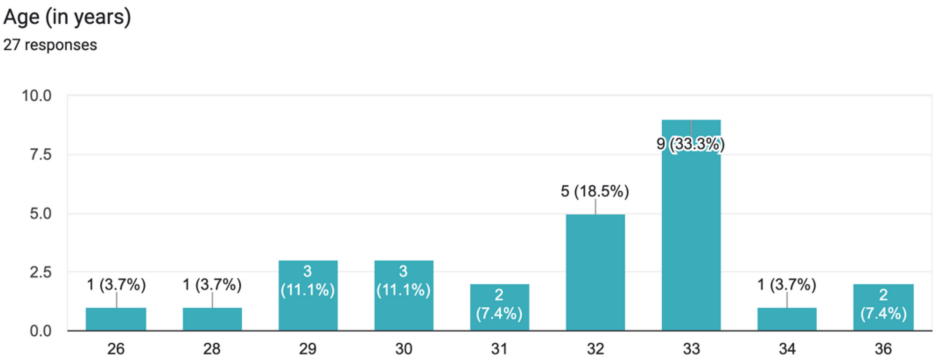


Figure 15. Age demographics of respondents.

5.4. Summary of Results on Explainability Levels

After conducting the survey, the Likert assessment was calculated using a straightforward method. First, the number of responses for each item was multiplied by the corresponding Likert scale value. Next, the result of this multiplication was divided by the total number of respondents. Figure 16 illustrates an example of this calculation.

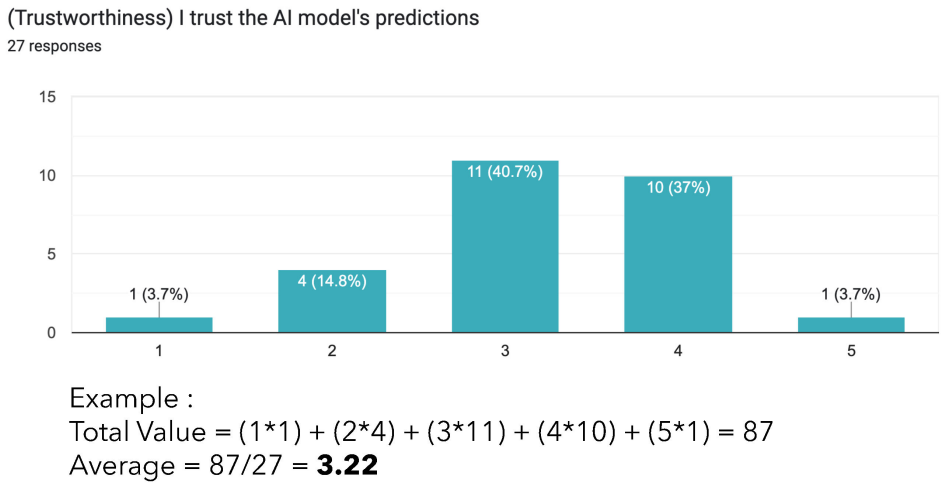


Figure 16. Example of Likert calculation.

Using this methodology, we evaluated the explainability of the five visualizations across the five metrics as summarized in Table 7.

Based on the results, several key highlights can be drawn. In general, the baseline model exhibits the lowest scores across all metrics. This model, which did not incorporate any XAI techniques, received particularly low ratings in causality and informativeness. This indicates that without the aid of explainability, users find it more challenging to trust or comprehend the model's decisions.

Table 7. Explainability level result.

Model	Trustworthiness	Causality	Transferability	Informativeness	Confidence
NoXAI	3.15	**2.85** [1]	3.18	**2.22** [1]	3.00
LIME	3.22	3.26	3.37	3.41	3.04
LIME + Text	3.41	3.52	3.48	3.67	3.30
SHAP	3.40	3.26	3.26	3.33	3.19
SHAP + Text [2]	**3.70**	**3.78**	**3.63**	**3.82**	**3.48**

[1] Both of these features received scores lower than 3 but are the most improved features after applying the XAI approach. [2] Model visualization using SHAP and generated text achieved the highest score across all features.

The SHAP model with text performed better across all five metrics, consistent with the findings from the preference question depicted in Figure 17. The reasons for this include its convincibility, ease of understanding, provision of more informative insights, and clearer explanations.

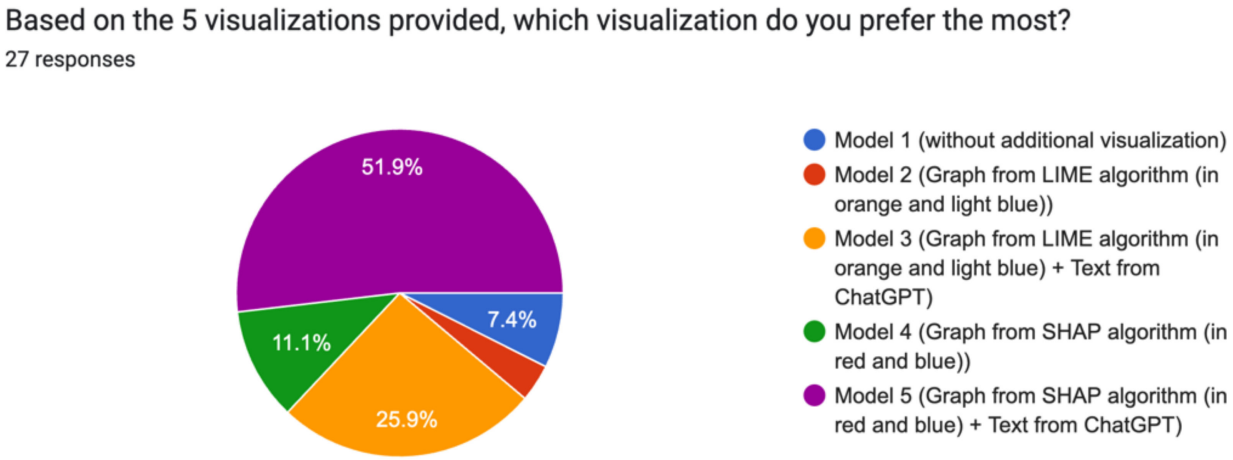


Figure 17. User visualization preference.

Furthermore, the inclusion of textual explanations—whether through LIME or SHAP—enhanced performance across all criteria. This suggests that users prefer models that offer detailed and comprehensible explanations. Notably, the largest improvements were observed in the metrics of causality and informativeness, highlighting that these factors are significantly impacted by the choice of explanation method.

In general, the levels of trustworthiness and confidence in the AI remained relatively stable. This indicates that respondents hold their inherent trust levels, yet they remain somewhat cautious regarding the AI predictions. The introduction of explainability resulted in only slight improvements in these perceptions.

In summary, the results suggest that incorporating XAI techniques, and also including textual components, significantly enhances user understanding and trust in AI models. This underscores the importance of providing clear and informative explanations, especially in areas critical for decision-making, such as causality and informativeness. The findings also point to the need for ongoing efforts to foster greater trust and confidence among users when interacting with AI systems.

6. Conclusions and Future Work

In this study, we propose a methodology for profiling high-risk patients with autosomal dominant polycystic kidney disease (ADPKD), in addition to traditional total kidney volume (TKV) calculations. By employing machine learning techniques on historical patient data, we were able to classify high-risk profiles. The dataset utilized for this research is authentic, having been collected over the past 70 years, but it required preprocessing prior to the modeling phase. This modeling process involved 63 experiments, integrating three datasets, utilizing seven machine learning algorithms, and applying three methods to address imbalanced data. The models were evaluated based on accuracy, precision, recall, F1 score, and AUC score, with the AUC score serving as the primary objective due to its robustness in handling imbalanced datasets.

Moreover, the interpretability of the prediction model for experts, particularly doctors, was a significant concern. Consequently, we implemented Explainable Artificial Intelligence (XAI) approaches. We selected LIME and SHAP for their capabilities in visualizing the reasoning behind model predictions by displaying feature importance and their contributions to the model prediction. Additionally, we explored the integration of generated text to further enhance user explainability.

Embedding a user-centered design process, we also conducted human evaluations to assess the explainability of the model visualizations. We proposed specific metrics to quantify explainability and surveyed participants to compare the five visualizations presented.

The findings from this survey are compelling, highlighting a significant increase in explainability levels when comparing the model without XAI techniques to those that implemented such methods. The inclusion of textual summarization notably improved user understanding. Future research should focus on fostering greater trust and confidence among users interacting with AI systems, enabling these systems to assist effectively in the decision-making process.

Author Contributions: L.D. contributed to the methodology, software development, and the writing of the original draft. H.N. provided supervision and contributed to the review and editing of the manuscript. N.H. also contributed to the review and editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: All doctors involved in this research provided their informed consent to participate in the study.

Data Availability Statement: The raw data used in this study are proprietary to the Polycystic Kidney Disease Outcomes Consortium (PKDOC). Access to the data was granted in accordance with the PKDOC Data Terms and Conditions. All data handling and analysis procedures strictly adhered to these terms to ensure compliance with the consortium's regulations and guidelines.

Acknowledgments: The first author would like to express the gratitude to Ministry of Education, Culture, Sports, Science and Technology (MEXT) Japan for providing the scholarship during the study at Kanazawa University, Japan.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Feature for Modeling Process

These are all the features used in the modeling process. The dataset containing all of these attributes is referred to as 'Set1'.

Table A1. Set1 feature.

Feature Type	Features	Values	Source
Demographic (5 features)	age, sex	continuous	raw data
	race, ethnic	categorical	raw data with null data handling
	flag_death	binary	raw data with null data handling
Intervention (4 features)	is_liver_procedure, is_cyst_procedure, is_kidney_procedure, is_blood_procedure	binary	performing one-hot-encoding from raw data
Substance Used (4 features)	is_consume_alcohol, is_consume_tobacco, is_consume_caffeine, is_consume_decaffeine	binary	performing one-hot-encoding from raw data
Hospitalization (1 feature)	is_hospitalization	binary	performing one-hot-encoding from raw data
Family history (8 feature)	is_affected_mother, is_affected_father, is_affected_siblings, is_affected_aunt_uncle, is_affected_grandparents, is_affected_son, is_affected_daughters, is_affected_others_family_member	binary	performing one-hot-encoding from raw data
CKD Stage Classification (1 feature)	CKD_classification	categorical	calculated and categorized from raw data
High Risk Profile Classification (1 feature)	is_high_risk_profile (Target Variable)	binary	calculated and categorized from raw data
Medical history and clinical record (77 features)	hypertension, flank_pain, back_pain, abdominal_pain, abdomen_protube, anorexia, nausea, anemia, headache, migraine, gout, fatique, changes_in_appetite, back_trouble, acne, asthma, hay_fever, rheumatic, pyelonephritis, rheumatic_heart, tuberculosis, jaundice, heart_attack, allergies, abmass, mood_changes, pneumonia, myeloma, cancer, scarlet_fever, constipation, diarrhea, insomnia, impotence, drowsiness, event_urinary_tract_infection, event_gross_hematuria, event_symptomatic_nephrolithiasis, event_edema, event_nocturia, event_non_exertional_chest_pain, event_shortness_of_breath_at_rest, event_exertional_chest_pain, event_inguinal_hernia, event_umbilical_hernia, event_intracranial_aneurysm, event_kidney_cyst_hemorrhage, event_asymptomatic_nephrolithiasis, event_cardiac_valve_disease, event_heart_palpitations, event_shortness_of_breath_with_exertion, event_diverticulosis, event_end_stage_renal_disease, event_cardiac_arrhythmia, event_shortness_of_breath, event_loss_of_consciousness, event_symptomatic_intracranial_aneurysm, coronary_artery_disease, event_left_ventricular_hypertrophys, event_diabetes_mellitus, event_seizures, event_severe_headache, event_diverticulitis, event_other_acute_neurological_event, event_kidney_cyst_infection, event_congenital_heart_disease, event_abdominal_mass, event_coronary_heart_failure, event_aneurysm, event_hepatic_venous_outflow_obstruction, event_ruptured_intracranial_aneurysm, event_ruptured_intracranial, event_inferio_vena_cava_compression, event_ascending_cholangitis, event_peripheralvascular_disease, event_kidney_stones, event_carotid_disease	binary	performing one-hot-encoding from raw data

Appendix B. Hyperparameter

These are the hyperparameters used in the training model. Most of them are the default values.

Table A2. Random Forest hyperparameters.

Parameter	Value
Number of Estimators	100
Max Depth	None
Min Samples Split	2
Min Samples Leaf	1
Max Features	auto
Bootstrap	True
Random State	42

Table A3. Logistic Regression hyperparameters.

Parameter	Value
Penalty	l2
C	1.0
Solver	lbfgs
Max Iterations	100
Multiclass	auto
Random State	None

Table A4. SVM hyperparameters.

Parameter	Value
C	1.0
Kernel	linear
Degree	3
Gamma	scale
Max Iterations	-1
Random State	None

Table A5. LightGBM hyperparameters.

Parameter	Value
Learning Rate	0.1
Number of Estimators	100
Max Depth	-1
Num Leaves	31
Subsample	1.0
Colsample By Tree	1.0
Min Child Weight	1×10^{-3}
Random State	42

Table A6. Gradient Boosting Tree hyperparameters.

Parameter	Value
Learning Rate	0.1
Number of Estimators	100
Max Depth	3
Min Samples Split	2
Min Samples Leaf	1
Subsample	1.0
Loss	deviance
Random State	42

Table A7. XGBoost hyperparameters.

Parameter	Value
Learning Rate (eta)	0.3
Max Depth	6
Min Child Weight	1
Subsample	1
Colsample by Tree	1
Number of Estimators	100
Gamma	0
Scale Pos Weight	1
Random State	42

Table A8. DNN hyperparameters.

Parameter	Value
Data Scaling	StandardScaler
Train–Test Split	Test Size: 0.2 Random State: 42
Input Layer	Shape (n_features)
Hidden Layer 1	Units 64, Activation ReLU
Hidden Layer 2	Units 32, Activation ReLU
Output Layer	Units 1, Activation sigmoid
Optimizer	Adam, Learning Rate 0.001
Loss Function	binary_crossentropy
Metrics	accuracy
Epochs	10
Batch Size	32
Validation Split	0.2
Early Stopping	Monitor: val_loss, Patience 5
Random State	42

Appendix C. Evaluation Metrics

This serves as the evaluation metric for the 63 experiments conducted.

Table A9. Result for the Evaluation Metrics.

Set Attribute	Algorithm	Imbalanced Method	Accuracy	Precision	Recall	F1	AUC
Set1	DNN	SMOTE	0.6864	0.6729	0.6864	0.6777	0.6126
Set1	DNN	ROS	0.6723	0.6458	0.6723	0.6519	0.5753
Set1	DNN		0.7119	0.6875	0.7119	0.6781	0.5968
Set1	GBT	SMOTE	0.6667	0.6282	0.6398	0.631	0.6397
Set1	GBT	ROS	0.678	0.6624	0.6871	0.6597	0.6871
Set1	GBT		0.7288	0.7054	0.6018	0.6017	0.6018
Set1	LightGBM	SMOTE	0.6949	0.6446	0.6432	0.6439	0.6432
Set1	LightGBM	ROS	0.6977	0.6525	0.6575	0.6547	0.6575
Set1	LightGBM		0.7288	0.6822	0.6385	0.6475	0.6385
Set1	LR	SMOTE	0.6808	0.6468	0.6623	0.65	0.6623
Set1	LR	ROS	0.6723	0.6417	0.6585	0.6439	0.6585
Set1	LR		0.7288	0.6895	0.619	0.6254	0.6189
Set1	RF	SMOTE	0.709	0.6592	0.6535	0.656	0.65348
Set1	RF	ROS	0.6893	0.6402	0.6415	0.6408	0.6415
Set1	RF		0.7232	0.678	0.6148	0.6207	0.6148
Set1	SVM	SMOTE	0.6582	0.6279	0.6434	0.6292	0.6434
Set1	SVM	ROS	0.661	0.6491	0.6723	0.6437	0.6723
Set1	SVM		0.7401	0.7277	0.6174	0.6218	0.6174
Set1	XGBoost	SMOTE	0.6751	0.6233	0.6239	0.6236	0.6239
Set1	XGBoost	ROS	0.6695	0.6235	0.6296	0.6258	0.6296
Set1	XGBoost		0.726	0.6785	0.6316	0.64	0.6316
Set2	DNN	SMOTE	0.7034	0.6999	0.7034	0.7015	0.6504
Set2	DNN	ROS	0.6356	0.6437	0.6356	0.6393	0.5878
Set2	DNN		0.7175	0.696	0.7175	0.6912	0.6132

Table A9. *Cont.*

Set Attribute	Algorithm	Imbalanced Method	Accuracy	Precision	Recall	F1	AUC
Set2	GBT	SMOTE	0.6667	0.6391	0.6569	0.6402	0.6569
Set2	GBT	ROS	0.6582	0.6434	0.6654	0.6392	0.6654
Set2	GBT		0.7316	0.7017	0.6137	0.6181	0.6137
Set2	LightGBM	SMOTE	0.6921	0.6473	0.6534	0.6498	0.6534
Set2	LightGBM	ROS	0.7006	0.6562	0.662	0.6587	0.662
Set2	LightGBM		0.726	0.6773	0.6389	0.6476	0.6389
Set2	LR	SMOTE	0.678	0.6491	0.6676	0.6512	0.6676
Set2	LR	ROS	0.6525	0.6323	0.6515	0.6302	0.6515
Set2	LR		0.7316	0.7017	0.6137	0.6181	0.6137
Set2	RF	SMOTE	0.6977	0.6475	0.6453	0.6463	0.6453
Set2	RF	ROS	0.7006	0.6495	0.6449	0.6469	0.6449
Set2	RF		0.7401	0.7216	0.6296	0.638	0.6296
Set2	SVM	SMOTE	0.661	0.6383	0.6577	0.6376	0.6577
Set2	SVM	ROS	0.6356	0.6367	0.6587	0.6229	0.6587
Set2	SVM		0.7147	0.6929	0.5695	0.5528	0.5695
Set2	XGBoost	SMOTE	0.6836	0.6314	0.6301	0.6307	0.6301
Set2	XGBoost	ROS	0.6977	0.6525	0.6575	0.6547	0.6575
Set2	XGBoost		0.7006	0.6383	0.6008	0.6052	0.6008
Set3	DNN	SMOTE	0.7203	0.7091	0.7203	0.7125	0.6519
Set3	DNN	ROS	0.732	0.717	0.732	0.719	0.689
Set3	DNN		0.7345	0.7202	0.7345	0.7017	0.6206
Set3	GBT	SMOTE	0.6695	0.653	0.6761	0.6503	0.676
Set3	GBT	ROS	0.6695	0.6604	0.6859	0.654	0.6858
Set3	GBT		0.7458	0.7363	0.6264	0.6334	0.6264
Set3	LightGBM	SMOTE	0.7034	0.6719	0.691	0.676	0.691
Set3	LightGBM	ROS	0.7006	0.667	0.684	0.6711	0.684
Set3	LightGBM		0.7345	0.6908	0.6451	0.6549	0.6451
Set3	LR	SMOTE	0.6838	0.6581	0.679	0.6596	0.679
Set3	LR	ROS	0.6751	0.6605	0.6851	0.6572	0.6851
Set3	LR		0.7288	0.6995	0.6067	0.6089	0.6067
Set3	RF	SMOTE	0.7062	0.668	0.6808	0.6722	0.6808
Set3	RF	ROS	0.6808	0.6412	0.6525	0.6446	0.6525
Set3	RF		0.7514	0.7193	0.6599	0.6722	0.6599
Set3	SVM	SMOTE	0.6808	0.6718	0.699	0.6658	0.699
Set3	SVM	ROS	0.6525	0.6624	0.6882	0.6433	0.6882
Set3	SVM		0.7147	0.6781	0.5793	0.5707	0.5793
Set3	XGBoost	SMOTE	0.7147	0.6825	0.7016	0.6872	0.7016
Set3	XGBoost	ROS	0.6723	0.6433	0.661	0.6451	0.661
Set3	XGBoost		0.7316	0.6857	0.6455	0.6548	0.6455
Set3	SVM		0.7147	0.6929	0.5695	0.5528	0.5695

Appendix D. Visualization of Extreme Cases

The extreme cases of the patient visualized with both LIME and SHAP.

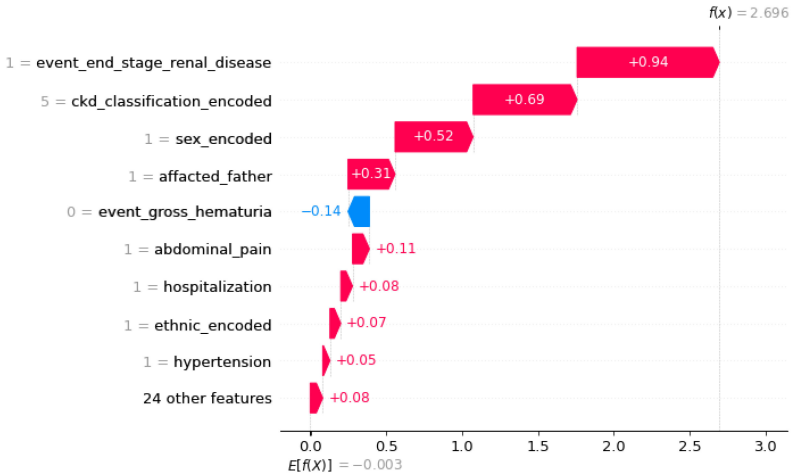


Figure A1. SHAP visualization for extreme case.

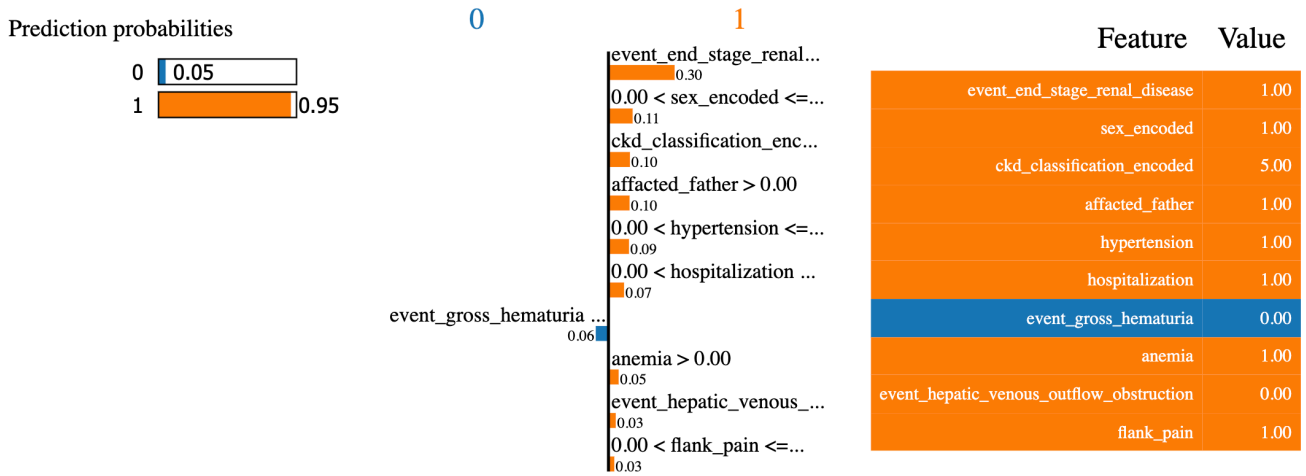


Figure A2. LIME visualization for extreme case.

References

1. Mahboob, M.; Rout, P.; Leslie, S.; Bokhari, S. *Autosomal Dominant Polycystic Kidney Disease*; StatPearls Publishing: Treasure Island, FL, USA, 2024; Volume 3.
2. Otsuka. What Is ADPKD, What Kind of Disease Is It. Available online: https://adpkd.jp/basic/about.html (accessed on 8 August 2024).
3. Willey, C.; Blais, J.; Hall, A.; Krasa, H.; Makin, A.; Czerwiec, F. Prevalence of autosomal dominant polycystic kidney disease in the European Union. *Nephrol. Dial. Transplant.* **2016**, *32*, 1356–1363. [CrossRef] [PubMed]
4. Chebib, F.; Torres, V. Autosomal dominant polycystic kidney disease: Core curriculum 2016. *Am. J. Kidney Dis.* **2016**, *67*, 792–810. [CrossRef] [PubMed]
5. Chapman, A.; Devuyst, O.; Eckardt, K.; Gansevoort, R.; Harris, T.; Horie, S.; Kasiske, B.; Odland, D.; Pei, Y.; Perrone, R.; et al. Autosomal-dominant polycystic kidney disease (ADPKD): Executive summary from a Kidney Disease: Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int.* **2015**, *88*, 17–27. [CrossRef]
6. Torres, V.; Chapman, A.; Devuyst, O.; Gansevoort, R.; Grantham, J.; Higashihara, E.; Perrone, R.; Krasa, H.; Ouyang, J.; Czerwiec, F.; et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N. Engl. J. Med.* **2012**, *367*, 2407–2418. [CrossRef] [PubMed]
7. Otsuka. Otsuka's JINARC the First-Ever Treatment Approved in Europe for Adults Living with ADPKD, a Chronic Genetic Kidney Disease. Available online: https://www.otsuka.co.jp/en/company/newsreleases/assets/pdf/20150528_1_01.pdf (accessed on 30 May 2024).
8. Foundation, P. Tolvaptan Treatment for ADPKD. Available online: https://pkdcure.org/tolvaptan/ (accessed on 30 May 2024).
9. Torres, V.; Chapman, A.; Devuyst, O.; Gansevoort, R.; Perrone, R.; Koch, G.; Ouyang, J.; McQuade, R.; Blais, J.; Czerwiec, F.; et al. Tolvaptan in later-stage autosomal dominant polycystic kidney disease. *N. Engl. J. Med.* **2017**, *377*, 1930–1942. [CrossRef]
10. Zhang, W.; Blumenfeld, J.; Prince, M. MRI in autosomal dominant polycystic kidney disease. *J. Magn. Reson. Imaging* **2019**, *50*, 41–51. [CrossRef]
11. JYNARQUE. Taking a Holistic Assessment Can Identify Appropriate Patients for JYNARQUE (Tolvaptan). Available online: https://www.jynarquehcp.com/identify-patients (accessed on 30 May 2024).
12. Taylor, J.; Thomas, R.; Metherall, P.; Gastel, M.; Cornec-Le Gall, E.; Caroli, A.; Furlano, M.; Demoulin, N.; Devuyst, O.; Winterbottom, J.; et al. An artificial intelligence generated automated algorithm to measure total kidney volume in ADPKD. *Kidney Int. Rep.* **2024**, *9*, 249–256. [CrossRef]
13. Ali, O.; Abdelbaki, W.; Shrestha, A.; Elbasi, E.; Alryalat, M.; Dwivedi, Y. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J. Innov. Knowl.* **2023**, *8*, 100333. [CrossRef]
14. Aljaaf, A.; Al-Jumeily, D.; Hussain, A.; Fergus, P.; Al-Jumaily, M.; Abdel-Aziz, K. Toward an optimal use of artificial intelligence techniques within a clinical decision support system. In Proceedings of the 2015 Science and Information Conference (SAI), London, UK, 28–30 July 2015.
15. Doniyorjon, M.; Madinakhon, R.; Shakhnoza, M.; Cho, Y. An Improved Method of Polyp Detection Using Custom YOLOv4-Tiny. *Appl. Sci.* **2022**, *12*, 10856. [CrossRef]
16. Mukhtorov, D.; Rakhmonova, M.; Muksimova, S.; Cho, Y. Endoscopic Image Classification Based on Explainable Deep Learning. *Sensors* **2023**, *23*, 3176. [CrossRef]
17. Bernardini, M.; Romeo, L.; Frontoni, E.; Amini, M. A Semi-Supervised Multi-Task Learning approach for predicting short-term kidney Disease evolution. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3983–3994. [CrossRef] [PubMed]

18. Almansour, N.; Syed, H.; Khayat, N.; Altheeb, R.; Juri, R.; Alhiyafi, J.; Alrashed, S.; Olatunji, S. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Comput. Biol. Med.* **2019**, *109*, 101–111. [CrossRef] [PubMed]
19. Raihan, M.J.; Khan, M.A.M.; Kee, S.H.; Nahid, A.A. Detection of the Chronic Kidney Disease Using XGBoost Classifier and Explaining the Influence of the Attributes on the Model Using SHAP. *Sci. Rep.* **2023**, *13*, 6263. [CrossRef] [PubMed]
20. Barredo Arrieta, A.; Diaz-Rodriguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
21. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
22. Lötsch, J.; Kringel, D.; Ultsch, A. Explainable artificial intelligence (XAI) in biomedicine: Making AI decisions trustworthy for physicians and patients. *BioMedInformatics* **2021**, *2*, 1–17. [CrossRef]
23. Gilpin, L.; Bau, D.; Yuan, B.; Bajwa, A.; Specter, M.; Kagal, L. Explaining explanations: An overview of interpretability of machine learning. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018.
24. Ribera, M.; Lapedriza, A. Can we do better explanations? A proposal of User-Centered Explainable AI. In Proceedings of the IUI Workshops '19, Los Angeles, CA, USA, 20 March 2019.
25. Schröer, C.; Kruse, F.; Gómez, J. A systematic literature review on applying CRISP-DM process model. *Procedia Comput. Sci.* **2021**, *181*, 526–534. [CrossRef]
26. Wirth, R.; Hipp, J. CRISP-DM: Towards a standard process model for data mining. In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, Manchester, UK, 11–13 April 2000.
27. Critical Path Institute. Polycystic Kidney Disease Outcomes Consortium. Available online: https://c-path.org/program/polycystic-kidney-disease-outcomes-consortium/ (accessed on 30 May 2024).
28. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model, Implementation Guide: Human Clinical Trials, Version 3.4 (Final). Available online: https://sastricks.com/cdisc/SDTMIG%20v3.4-FINAL_2022-07-21.pdf (accessed on 29 November 2023).
29. Jo, W.; Kim, S.; Kim, K.; Suh, C.; Kim, J.; Kim, H.; Lee, J.; Oh, W.; Choi, S.; Pyo, J. Correlations between renal function and the total kidney volume measured on imaging for autosomal dominant polycystic kidney disease: A systematic review and meta-analysis. *Eur. J. Radiol.* **2017**, *95*, 56–65. [CrossRef]
30. Park, S.; Jeong, T. Estimated glomerular filtration rates show minor but significant differences between the single and subgroup creatinine-based Chronic Kidney Disease Epidemiology Collaboration equations. *Ann. Lab. Med.* **2019**, *39*, 205–208. [CrossRef]
31. American Kidney Fund. Stages of Kidney Disease (CKD). Available online: https://www.kidneyfund.org/all-about-kidneys/stages-kidney-disease (accessed on 30 May 2024).
32. Irazabal, M.; Rangel, L.; Bergstralh, E.; Osborn, S.; Harmon, A.; Sundsbak, J.; Bae, K.; Chapman, A.; Grantham, J.; Mrug, M.; et al. Imaging classification of autosomal dominant polycystic kidney disease: A simple model for selecting patients for clinical trials. *J. Am. Soc. Nephrol.* **2015**, *26*, 160–172. [CrossRef]
33. Rastogi, A.; Ameen, K.; Al-Baghdadi, M.; Shaffer, K.; Nobakht, N.; Kamgar, M.; Lerma, E. Autosomal dominant polycystic kidney disease: Updated perspectives. *Ther. Clin. Risk Manag.* **2019**, *15*, 1041–1052. [CrossRef] [PubMed]
34. Franklin, G.; Stephens, R.; Piracha, M.; Tiosano, S.; Lehouillier, F.; Koppel, R.; Elkin, P. The Sociodemographic Biases in Machine Learning Algorithms: A Biomedical Informatics Perspective. *Life* **2024**, *14*, 652. [CrossRef] [PubMed]
35. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
36. Japkowicz, N.; Stephen, S. The class imbalance problem: A systematic study. *Intell. Data Anal.* **2002**, *6*, 429–449. [CrossRef]
37. He, H.; Garcia, E. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
38. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Panda, N. A review on logistic regression in medical research. *Natl. J. Community Med.* **2022**, *13*, 265–270. [CrossRef]
40. Burges, C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
41. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; NeurIPS: San Diego, CA, USA, 2017.
42. Krauss, C.; Do, X.; Huck, N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the SP 500. *Eur. J. Oper. Res.* **2017**, *259*, 689–702. [CrossRef]
43. Chen, T.; Guestrin, C. XGBoost. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016.
44. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
45. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
46. Zhou, Z. *Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2021.
47. Nai, Y.; Teo, B.; Tan, N.; O'Doherty, S.; Stephenson, M.; Thian, Y.; Chiong, E.; Reilhac, A. Comparison of metrics for the evaluation of medical segmentations using prostate MRI dataset. *Comput. Biol. Med.* **2021**, *134*, 104497. [CrossRef] [PubMed]

48. Müller, D.; Soto-Rey, I.; Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **2022**, *15*, 210. [CrossRef] [PubMed]
49. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Med. Imaging* **2015**, *15*, 29. [CrossRef] [PubMed]
50. Lipton, Z.; Steinhardt, J. Why do tree based models still outperform deep learning on tabular data? *arXiv* **2019**, arXiv:2207.08815. [CrossRef]
51. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. *arXiv* **2019**, arXiv:1907.10902.
52. Ribeiro, M.; Singh, S.; Guestrin, C. Model-Agnostic Interpretability of Machine Learning. *arXiv* **2016**, arXiv:1606.05386.
53. Ribeiro, M.; Singh, S.; Guestrin, C. "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, San Diego, CA, USA, 12–17 June 2016.
54. Lundberg, S.; Lee, S. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874.
55. Molnar, C. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, BC, Canada, 2020.
56. Doshi-Velez, F.; Kim, B. Considerations for Evaluation and Generalization in Interpretable Machine Learning. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*; Springer: Cham, Switzerland, 2018; pp. 3–17.
57. Joshi, A.; Kale, S.; Chandel, S.; Pal, D. Likert scale: Explored and explained. *Br. J. Appl. Sci. Technol.* **2015**, *7*, 396–403. [CrossRef]