PLOS ONE

RESEARCH ARTICLE

A machine learning-based predictive model of causality in orthopaedic medical malpractice cases in China

Qingxin Yang[1], Li Luo[1], Zhangpeng Lin[1], Wei Wen[1], Wenbo Zeng[2], Hong Deng[1]*

1 School of Forensic Medicine, Kunming Medical University, Kunming, China, **2** West China Hospital of Sichuan University, Chengdu, China

* Dhong23@163.com

Citation: Yang Q, Luo L, Lin Z, Wen W, Zeng W, Deng H (2024) A machine learning-based predictive model of causality in orthopaedic medical malpractice cases in China. PLoS ONE 19(4): e0300662. https://doi.org/10.1371/journal.pone.0300662

Editor: Roberto Scendoni, University of Macerata: Universita degli Studi di Macerata, ITALY

Received: August 3, 2023

Accepted: February 27, 2024

Published: April 17, 2024

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Purpose

To explore the feasibility and validity of machine learning models in determining causality in medical malpractice cases and to try to increase the scientificity and reliability of identification opinions.

Methods

We collected 13,245 written judgments from PKULAW.COM, a public database. 963 cases were included after the initial screening. 21 medical and ten patient factors were selected as characteristic variables by summarising previous literature and cases. Random Forest, eXtreme Gradient Boosting (XGBoost) and Light Gradient Boosting Machine (LightGBM) were used to establish prediction models of causality for the two data sets, respectively. Finally, the optimal model is obtained by hyperparameter tuning of the six models.

Results

We built three real data set models and three virtual data set models by three algorithms, and their confusion matrices differed. XGBoost performed best in the real data set, with a model accuracy of 66%. In the virtual data set, the performance of XGBoost and LightGBM was basically the same, and the model accuracy rate was 80%. The overall accuracy of external verification was 72.7%.

Conclusions

The optimal model of this study is expected to predict the causality accurately.

1. Introduction

Machine learning (ML) is subfield of artificial intelligence (AI) that focuses on teaching computers to identify and interpret patterns within data through training [1]. ML have demonstrated potential across various domains within the biomedical sciences, such as genomics [2, 3], clinical medicine [4, 5] and forensic medicine [6, 7]. Models that have been published in clinical medicine can enhance the alertness of clinicians, carry out diagnostic procedures, predict events pertinent to clinical practice, and steer the process of making clinical decisions [8, 9]. However, few models have been used in clinical practice, which may be due to the challenges of machine learning models in feature selection, model complexity and generalization ability, the quantity and quality of training data, model interpretability, and other ethical and legal factors [10]. Within the realm of forensic medicine, the implementation of AI has the potential to augment the capabilities of human experts, effectively mitigating the inherent subjectivity and bias associated with conventional forensic methodologies. Forensic anthropology primarily involves the reconstruction of biological profiles for deceased individuals to ascertain their identity, including attributes such as sex, age at the time of death, and ancestry [11, 12]. In forensic odontology, AI has been utilized to forecast age and gender from dental characteristics, facilitating both human identification processes and the analysis of bite marks [13]. In disability assessment, researchers have tried to combine ML with the International Classification of Functioning, Disability, and Health (ICF) to assess the degree of disability more accurately and conveniently [14, 15]. In addition, ML also has some applications in forensic pathology, forensic genetics and other forensic branches. But it is scarcely any used in medical malpractice.

In recent years, the number of medical malpractice tort liability cases in China has increased, and orthopaedics is one of the departments with the most cases [16, 17]. Due to medical disciplines' high professionalism and complexity, judges need to rely on professional technical assistance to adjudicate such cases. In order to ensure fairness and justice of verdicts, China has established a "dual-mode" structure of two third-party authentication organisations, the medical association identification and the judicial appraisal institution, which is similar to the system of single joint expert (SJE) [18]. The medical association identification and the judicial appraisal institution do not represent any of the parties and have a neutral status. They accept the commission of the court by experts to conduct a retrospective analysis of the cases and then issue a written appraisal opinion which will become scientific evidence [19].

The identification of medical malpractice cases in China needs to consider four issues: (1) Whether there is any fault in the medical treatment process; (2) Whether the doctor has caused substantial harm to the patient; (3) Whether there is a factual causality between the physician's fault and the patient's damage; (4) The degree of the factual causality. Experts can make scientific identifications based on textbooks, clinical guidelines and etc., to determine whether there is any fault in the medical treatment process. However, as for the degree of factual causality, how to judge it scientifically is still the slip of everyone's debate. Scholars have put forward many different theories, such as "Bolam standard", "Forcier-Lacerte medicolegal causal analysis model" [20], "Integration of Forensic Epidemiology and the Rigorous Evaluation of Causation Elements (INFERENCE)" [21]. Many causal analysis methods are being applied in various countries, and experts inevitably mix subjective factors in practical application. Especially in the face of complex causal issues, the opinions of different experts in the same case may be different [22]. This makes the objectivity and reliability of the identification opinions questionable [23]. Therefore, improving the reliability of identification opinions and reducing subjective factors has become an urgent issue that needs to be solved.

In China, the conventional procedure for medical malpractice tort liability cases typically unfolds in a structured sequence: Initially, a dispute arises between the patient and the hospital,

prompting both parties to jointly initiate legal proceedings in court. Subsequently, the court entrusts the case to a medical association, composed of clinical medicine experts, or a judicial appraisal institution staffed with forensic experts for an impartial evaluation. These experts meticulously examine the case and formulate their identification opinions, which are then submitted to the judicial technician for thorough review. The technician's assessment is ultimately conveyed back to the judge, who renders a verdict based on the expert findings and the merits of the case. Even though the forensic expert is highly trained and uses specific guidelines, subjectivity can lead to inaccuracies in the evaluations, for example, errors due to incorrect analysis of the data available or the methodology followed in complex cases. Therefore, we present a new process that shows how machine learning can be integrated into decision-making process to reduce the margin of error in evaluation (Fig 1). We commence by leveraging a dataset comprising medical malpractice cases within the field of orthopedics in China as our foundational training set. We meticulously curate medical and patient factors as key features and harness the power of machine learning to construct an optimal causality prediction model. Building upon this traditional methodology, we introduce an innovative process: upon the court's delegation of a case to expert evaluators, the case details are meticulously processed through feature extraction and fed into our premier machine learning model. Subsequently, the judicial technician meticulously compares the model's classification outcomes with the expert's assessments. Should the findings align, the expert's opinion is deemed highly reliable, thereby informing the judge's decision-making process. Conversely, in the event of discordant conclusions, the court must engage a new expert to scrutinize the inconsistencies and provide a cogent explanation, ensuring the integrity of the judicial process.

In this way, our purpose is to propose a ML supplementary means based on real case data, aiming to explore the feasibility and validity of machine learning models in determining causality in medical malpractice cases and to try to increase the scientificity and reliability of identification opinions. as well as ultimately improve the fairness of court verdicts. The complete flow chart of this study is shown in Fig 2.

2. Material and methods

2.1 Source of data

Firstly, it was searched in the PKULAW.COM (a public database) under the following conditions: (1) Cause of action: "medical malpractice tort liability"; (2) Instrument type: "Judgment"; (3) 2010.1.1–2023.5.1; (4) Full-text search field: "fracture" and "judicial appraisal". A total of 13,245 cases were retrieved. **Secondly**, the following additional criteria were included: (1) The cause of the malpractice cases was orthopaedic-related diseases; (2) The judicial appraisal institutions issued the identification opinion; (3) The judgment contained an analytical explanation for determining the degree of the causality; (4) The causality was clearly stated in the identification opinions; (5) Only one of the hospitals involved was at fault; (6) The full name of the hospital was included. We filtered the case contents one by one according to condition. 1024 cases were included, while 12235 cases were excluded. **Finally**, based on whether the identification opinions were admissible in the final judgment, 963 cases were included, while 61 cases were excluded from the total of 1024 cases. All the data in this study are from a public database, which does not contain any privacy-related information after processing, so there is no need for an ethical review.

2.2 Feature selection and data preprocessing

The label was "Causality". Guidance for judicial expertise of medical malpractice (China SF/T 0097–2021) [24] classified causality into six degrees: (1) No causality: The consequences of

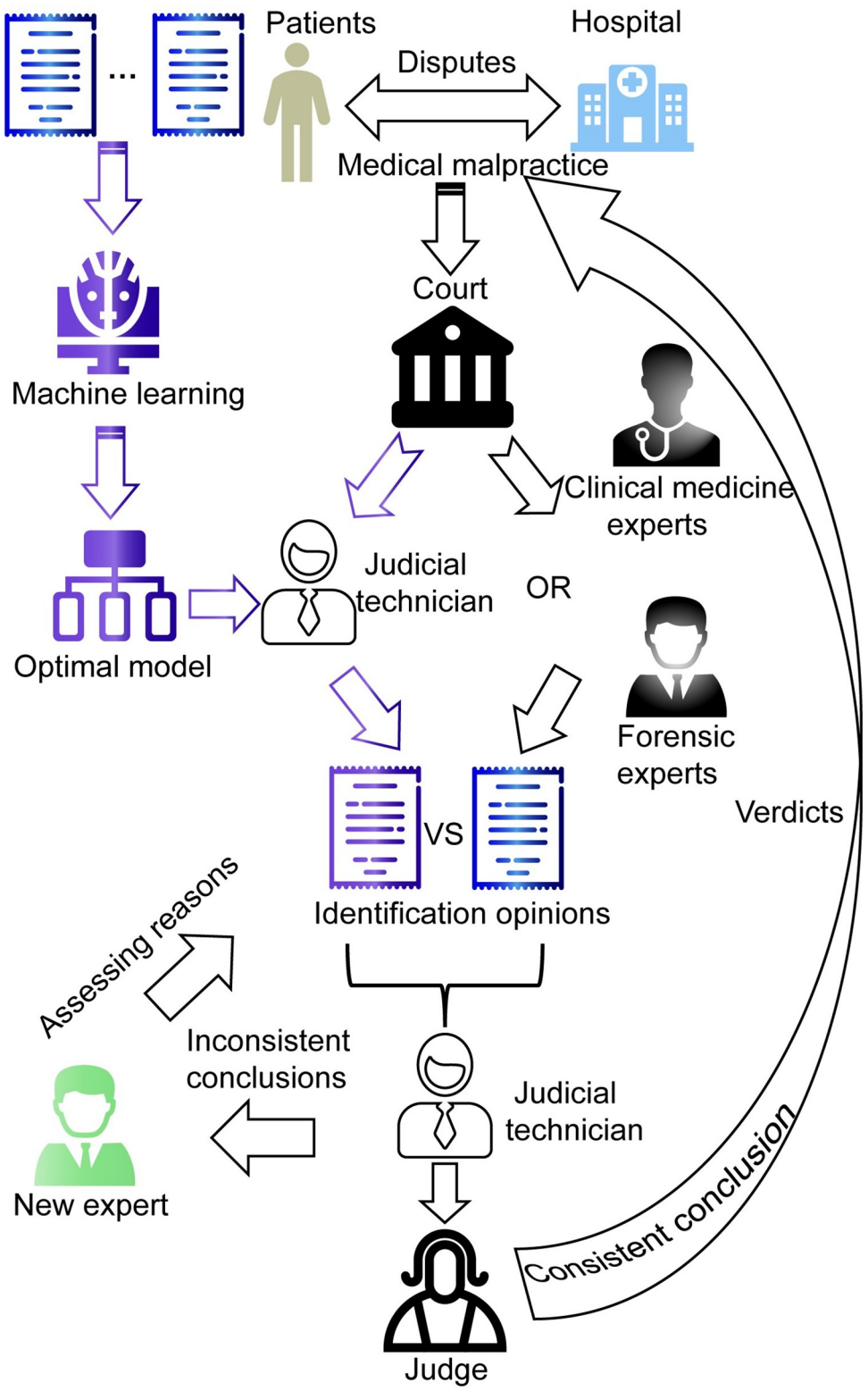

Fig 1. The new decision-making process of medical malpractice.

https://doi.org/10.1371/journal.pone.0300662.g001

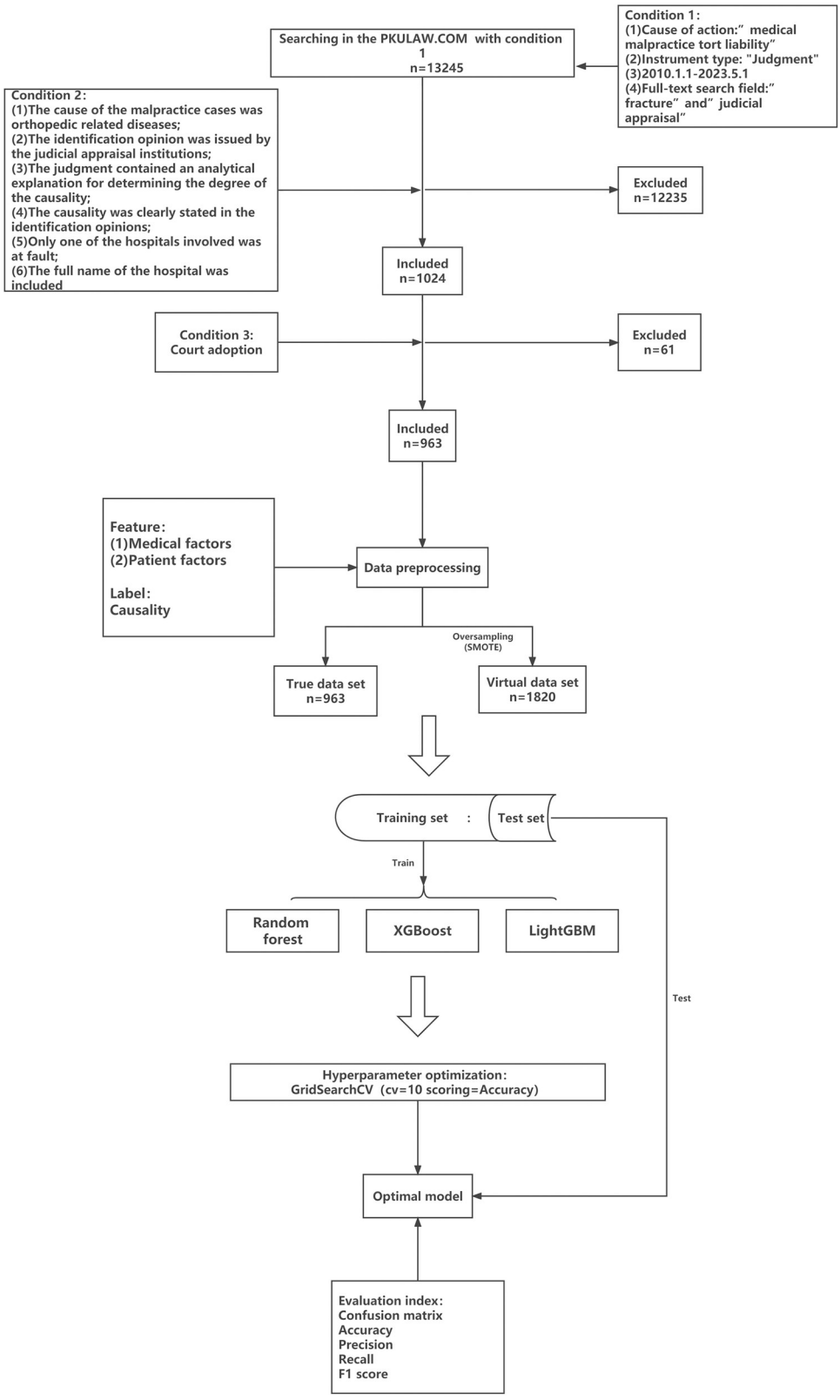

Fig 2. The complete flow chart of this study.

https://doi.org/10.1371/journal.pone.0300662.g002

damage were almost entirely due to patient factors, and there is no essential correlation with medical behaviour. (2) Minor causality: most of the damage consequences were due to patient factors, and the medical factors induced or slightly promoted and aggravated the effects. (3) Secondary causality: The damage consequences were primarily due to patient factors, and the medical factors played a role in promoting and aggravating. (4) Equal causality: Medical factors and patient factors played similar roles in forming damage consequences, and it was difficult to distinguish the primary and secondary. By parity of reasoning, there was the main and whole causality.

Thabet et al. 's [25] meta-analysis of orthopaedic litigation divided the factors causing litigation into medical factors (diagnostic faults and procedural faults) and patient factors (nature and location of injury). Dong et al. 's [26] graph theory analysis study established a complex network of medical malpractice in China, in which factors such as the technical and non-technical faults of the medical provider, the type of disease of the patient, and the degree of damage caused by the medical provider to the patient have their respective proportions.

We summarized the previous literature reports [14, 25–31], after fully understanding those common factors and considering the reality of medical malpractice cases in China, and divided the influencing factors on the degree of causality into medical factors (91 technical faults and 29 non-technical faults) and 10 patient factors (S1 Table). Regarding the attribution of medical malpractice, it is generally observed that an increased number of affirmative responses in medical factors correlates with a higher degree of hospital liability in the case. Conversely, a greater number of affirmative responses in patient factors typically diminish the proportion of medical responsibility. However, exceptions to this general rule may arise when the hospital's negligence results in exceptionally severe consequences, or when the physician's error, though entirely preventable, was inescapable. To navigate these complexities, we employed a hybrid approach, integrating data-driven insights with domain expertise. This methodology involved extensive consultation with five forensic and five clinical experts, culminating in the identification of a refined set of 31 characteristics for further analysis. **Medical factors**: (1) Hospital level; (2) Missed diagnosis and delayed treatment; (3) Inadequate preoperative preparation; (4) Insufficiency of operative pointer; (5) Inadequate therapeutic schedule; (6) Inadequate alternative treatment; (7) Inadequate operation technique; (8) Inadequate manual reduction; (9) Inadequate external fixation; (10) Inadequate internal fixation; (11) Anesthesia problem; (12) Inadequate nursing and observation; (13) Inadequate postoperative examination; (14) Inadequate medication use; (15) Insufficient recognition; (16) Inadequate hospital management; (17) Inadequate discharge instructions; (18) Inadequate contingency handling; (19) Inadequate consultation and referral; (20) Inadequate informed consent; (21) Medical record problem. **Patient factors**: (1) Over 60 years old; (2) Traumatic or not; (3) Number of other diseases (Such as diabetes, osteoporosis, nutritional status, etc.); (4) Comminuted fracture or not; (5) Number of hospitalisations; (6) Damage consequence (Divided into: no, prolonged course of disease, aggravated disease, disability, death); (7) Lack of compliance; (8) Severe illness or progress rapidly; (9) Uncommon disease; (10) Prognosis of disease. Crafting an effective dataset necessitates the meticulous identification of a cadre of salient, quantifiable factors and the development of unbiased scales that enable each unique case to be distilled into a standardized array of descriptors. Given the intricate interplay of medical and legal expertise inherent in Chinese medical malpractice adjudications, conventional natural language processing (NLP) techniques fall short in capturing the pertinent details. Consequently, we have adopted a meticulous manual reading approach to feature extraction, ensuring the fidelity and nuance of the data are preserved. Some features were grade variables, while the rest were binary variables. The above features and label assignments are shown in Table 1. By converting the information in the judgment into characters, and then calculating the common features by the ML algorithm, it is presented as a prediction model for the causality.

Table 1. Features and label assignment.

Name	Assignments
Causality	No = 0; Minor = 1; Secondary = 2; Equal = 3; Main = 4; Whole = 5;
Damage consequence	No = 0; Prolonged Course of disease = 1; Aggravated disease = 2; Disability = 3; Death = 4;
Hospital level	No = 0; First = 1; Second = 2; Third = 3;
Number of other diseases	No = 0; One = 1; Two = 2; Three or more = 3;
Number of hospitalisations (median = 2)	Less than or equal to 2 = 0; Greater than 2 = 1;
Medical factors Patient factors	No = 0; Yes = 1; No = 0; Yes = 1;

https://doi.org/10.1371/journal.pone.0300662.t001

2.3 Model selection and establishment

The 31 feature variables and one label selected in the true data set of this study were all classified variables after preprocessing, and the categories of labels, as shown in Fig 3, belong to the labelled imbalanced data set. Our goal was to train the model based on the training set and then accurately classify and predict the test set based on the model. Therefore, this paper chose three Ensemble Learning models based on the Decision tree model in machine learning classification algorithm: Random Forest, XGBoost, and LightGBM [32]. Decision Tree was a flow-chart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. These three models can sample according to weights to balance the test set data, perform well for imbalanced multi-classification problems, and output the importance of features, which has guiding significance for the subsequent feature selection. In addition, machine learning algorithms based on decision tree-based models also have the advantages of less data preparation, no data normalization, no data scaling, and missing values do not affect the modeling process. Last but not least, the tree model is highly interpretable, and the black-box problem was solved by output tree structure.

Python 3.10.7; Compiler: Jupyter Notebook were used as the language environment of this study. First, we imported the python base function library including "numpy", "scipy",

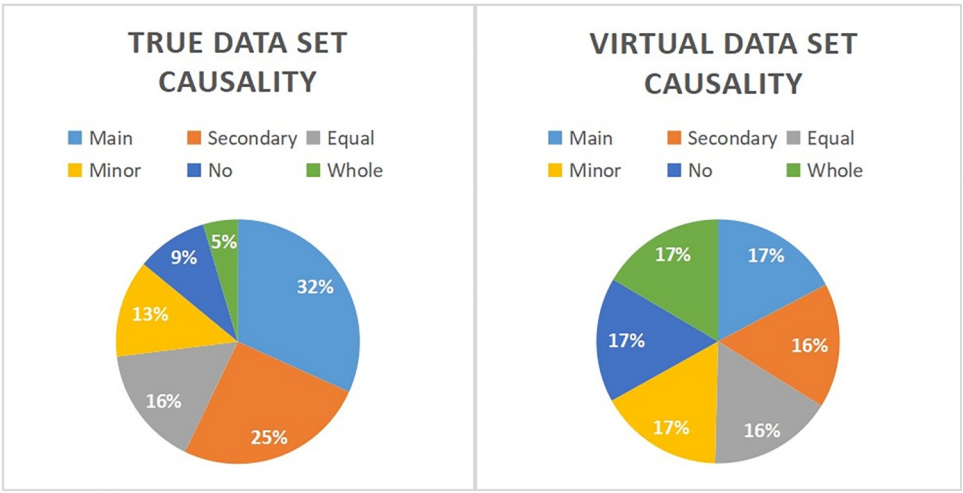

Fig 3. Label "Causality" proportion of each category.

https://doi.org/10.1371/journal.pone.0300662.g003

"pandas"and the drawing function library containing "matplotlib.pyplot", "seaborn". Then, we divided the real data set into the training and test sets with a weight ratio of 7:3, "random_state" = 6, using train_test_split from sklearn.model_selection. We calculated Random Forest using RandomForestClassifier from sklearn.ensemble, XGBoost using XGBClassifier from xgboost.sklearn and LightGBM using LGBMClassifier from lightgbm.sklearn. After training the three models with the training set, the Grid Search Cross Validation (GridSearchCV, cv = 10, scoring = accuracy) was used to optimise the model hyperparameters. As for Random Forest, its hyperparameters had "n_estimators: range (10, 300, 10)", "min_samples_split: range (5, 50, 5)", "min_samples_leaf: range (2, 40, 2)", "max_depth: range (1, 30, 2)", "criterion: 'gini', 'entropy' ", "class_weight: None, 'balanced' ". As for XGBoost, "max_depth: [3, 5, 7] ", "learning_rate: [0.1, 0.01, 0.001] ", "subsample: [0.1, 0.01, 0.001] ", "colsample_bytree: [0.5, 0.7, 1] ", "gamma: [0, 0.1, 0.2, 0.3, 0.4] ", "reg_alpha: [0, 0.001, 0.005, 0.01, 0.05] ", "reg_lambda: [0, 0.001, 0.005, 0.01, 0.05] ". The hyperparameters of LightGBM was same as XGBoost except "gamma". GridSearchCV exhaustively searches for all possible combinations in a given parameter space, then evaluates the performance of each combination, and finally selects the parameter combination with the best performance. Finally, the real optimal model was obtained. At the same time, to form a virtual data set with a labelled ratio of 1:1:1:1:1:1 (Fig 3), the Synthetic Minority Oversampling Technique (SMOTE), an oversampling method from imblearn.over_sampling, was used to expand the true data set. The virtual data set was divided into the training set, and the test set with 9:1 [32], "random_state" = 6, and then the virtual optimal model was established according to the above method.

2.4 Model performance evaluation

For multiple classification problems, the confusion matrix and the following four indicators were adopted as the criteria to measure the overall performance of the model:(1) Accuracy; (2) Precision; (3) Recall; (4) F1 score. True Positive (TP): A positive example of being correctly predicted. That is, the true value of the data was a positive example, and the predicted value was also a positive example. True Negative (TN): Counter-examples that their true data value was a counter-example, and the predicted value was also a counter-example. False Positive (FP): Positive example of misprediction. That is, the true value of the data was a negative example, but it was incorrectly predicted to be a positive example. False Negative (FN): A counter-example of being incorrectly predicted, in which the true value of the data was a positive example but incorrectly predicted to be a negative example [33].

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

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

$$Recall = \frac{TP}{TP + FN}$$

2.5 Feature importance ranking

We used the attribute "feature_importances_" to see the importance of the feature.

2.6 External data validation

We collected 11 orthopaedic medical malpractice cases from a judicial expertise centre in Yunnan Province in 2021–2022 as an external data set to verify the performance of the best model.

3. Results

The 963 cases included in this study involved several hospitals and judicial appraisal institutions in 29 provinces all over China. There are 21 factors of doctors and ten factors of patients.

3.1 True data set

The true data set was divided into the training set and the test set according to 7:3. That is, 289 cases were taken as the test set and put into the optimal prediction model of the three algorithms, respectively. The confusion matrix is shown in Fig 4. Among the three models, the precision of label classification as "No" is the highest, reaching 90%; The precision of "Main" is higher, reaching 70%-75%; The precision of "Minor", "Secondary", and "Equal" is slightly

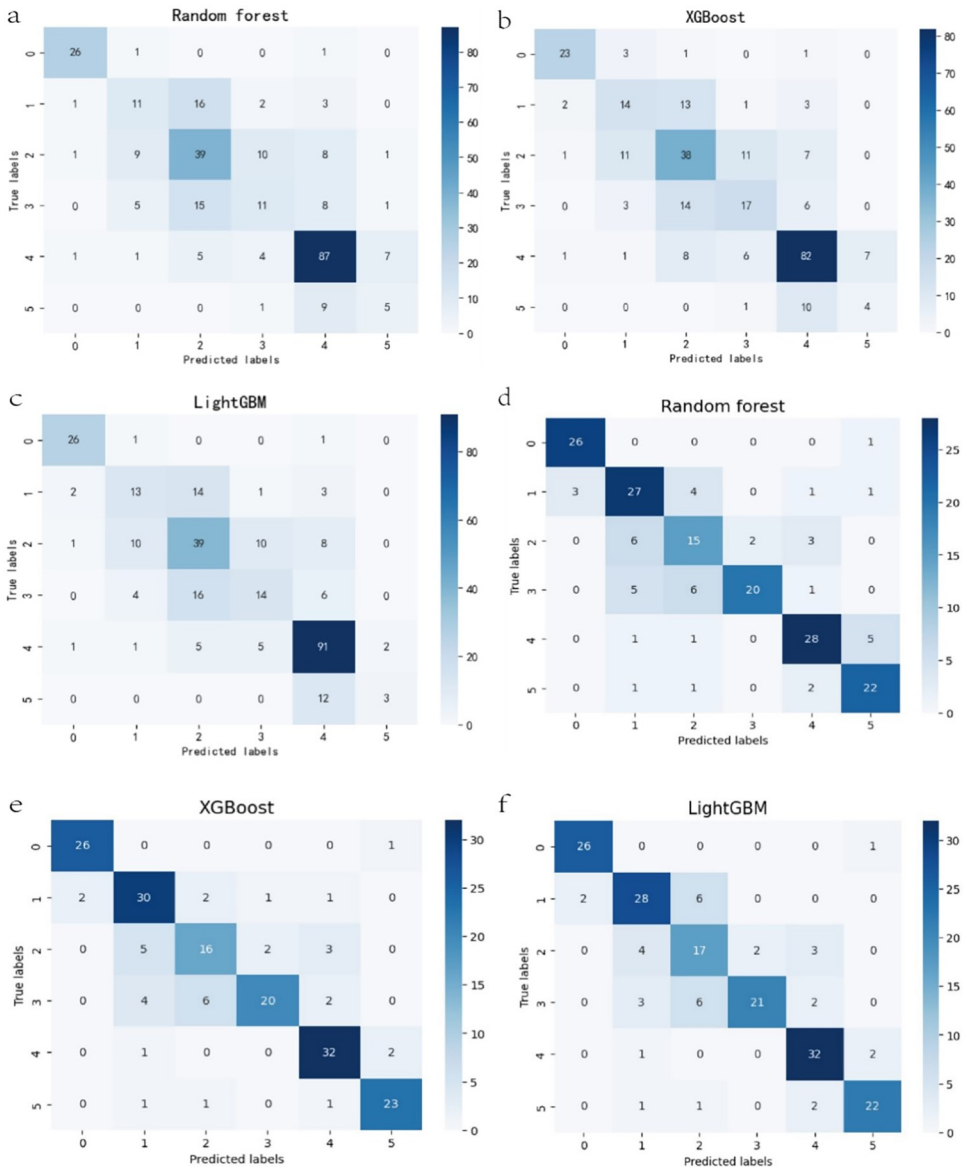

Fig 4. (a, b, c) The test set confusion matrix of the true data set. (d, e, f) The test set confusion matrix of the virtual data set. ("Predicted Label": No = 0; Minor = 1; Secondary = 2; Equal = 3; Main = 4; Whole = 5).

https://doi.org/10.1371/journal.pone.0300662.g004

worse, at 40–55%; The precision of "Whole" is only about 40%, which is related to the small number of "Whole" in the true data set.

3.2 Virtual data set

The virtual data set of n = 1280 was formed by oversampling with SMOTE and divided into training sets and test sets according to 9:1. That is, 128 cases were put into the optimal prediction model of the three algorithms, respectively. The confusion matrix is shown in Fig 4. Among the three models, the precision of "No" and "Main" was the highest, which can reach more than 90%. The precision of "Minor", "Equal", and "Whole" was higher, reaching 70%-85%. "Secondary" precision was the worst, at about 55%.

3.3 Model overall performance comparison

Accuracy, Precision, Recall and F1 score were used to evaluate the model's overall performance. The higher the value, the better the model performance. The comparison shows that the XGBoost model performed best in the real data set (Table 2). The performance of XGBoost and LightGBM in virtual data sets was better than that of Random Forest. In addition, the same algorithm performs better in virtual data set than in true data set. However, because the virtual data set was formed by oversampling, there may be problems with overfitting and data leakage.

3.4 External data validation

Eleven orthopaedic medical malpractice cases from a judicial expertise centre in Yunnan, China, were preprocessed and imported into an optimal model trained from a virtual dataset using XGBoost. The overall accuracy of the model was 72.7%. The results are shown in Table 3.

3.5 Model interpretability and feature importance

ML algorithms contend with the issue of opacity, where the system fails to offer any coherent rationale or satisfactory elucidation for its decisions, a conundrum often referred to as "the black-box dilemma." The enigmatic character of ML algorithms poses a significant comprehension challenge for human understanding [34, 35]. Entrusting crucial decisions to a black-box model created a necessary need for ML algorithms to be explainable for their decision-

Table 2. The performance of three models in true and virtual data sets. (RF: Random Forest; XGB: eXtreme Gradient Boosting; LGBM: Light Gradient Boosting Machine.).

	The true data set			**The virtual data set**		
	RF	XGB	LGBM	RF	XGB	LGBM
Accuracy	0.62	0.66	0.64	0.76	0.80	0.81
Precision	0.60	0.64	0.63	0.77	0.80	0.80
Recall	0.62	0.66	0.64	0.76	0.80	0.80
F1 score	0.61	0.63	0.63	0.76	0.80	0.80

https://doi.org/10.1371/journal.pone.0300662.t002

Table 3. Comparison of external validation results.

True value	0	3	1	4	4	3	3	1	4	2	5
Predicted value	0	3	2	4	4	2	3	1	4	2	4

https://doi.org/10.1371/journal.pone.0300662.t003

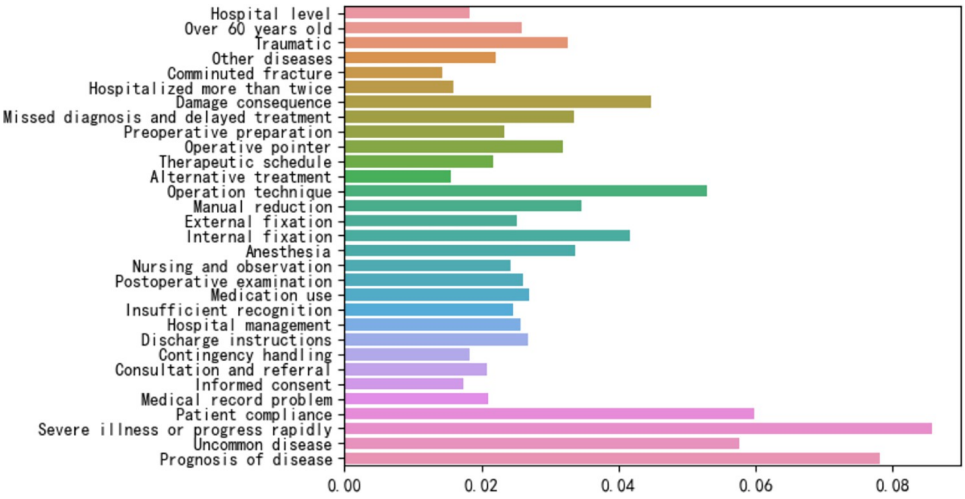

Fig 5. The feature importance of XGBoost, the optimal model for true data set.

https://doi.org/10.1371/journal.pone.0300662.g005

making process [36]. Therefore, we chose the ML based decision trees which solved the "black box" problem by integrating the prediction results of multiple decision trees. Each decision tree was a "white box" that could be understood and interpreted. When these decision trees were integrated together, random forests could provide more stable and reliable predictions while maintaining the explanatory properties of individual decision trees. The model is capable of assigning an importance score to each feature, which serves as a valuable tool for elucidating the intricacies of the model's predictive process. Furthermore, random forests are adept at uncovering interactions and nonlinear dynamics between features, thereby offering a more holistic and nuanced interpretation of the underlying data relationships. But the ML based decision trees construction process was done automatically, and the model learns and generates decision trees on its own based on the provided data, so there was usually no need for additional human opinion or intervention. The quality of the model depended on feature selection and hyperparameter optimization which can be manually intervened.

Knowing the feature importance to the prediction model helped us better understand the decisions and actions of the model [37]. The feature importance of XGBoost in the optimal model of the true data set is shown in Fig 5. The top 10 features are: (1) Severe illness or progress rapidly; (2) Prognosis of disease; (3) Lack of compliance; (4) Uncommon disease; (5) Inadequate operation technique; (6) Damage consequence; (7) Inadequate internal fixation; (8) Inadequate manual reduction;(9) Anesthesia problem; (10) Missed diagnosis and delayed treatment. Among the three models, five items were ranked in the top ten: (1) Severe illness or progress rapidly; (2) Prognosis of disease; (3) Inadequate operation technique; (4) Damage consequence; (5) Inadequate internal fixation.

4. Discussion

In the true data set of this study, the distribution of labelled "Causality" was similar to the normal distribution, and relatively few cases were classified as "No" and "Whole". There were a few cases in which the court ruled that there was no causality or whole causality between medical factors and patient damage consequences. The data set's imbalance nature led to a decline in the accuracy of machine learning models, with the optimal model only reaching 66%. However, when we used the oversampling method to form a virtual balanced data set, the model's accuracy increased significantly to 80%. To verify the real performance of the model, we used

11 cases as an external dataset for external validation, and its overall accuracy was 72.7%, close to the ideal 80%. This data set only collected cases of orthopaedic-related medical malpractice, and the model can only be used in orthopaedic departments. However, the accuracy of 72.7% indicates that the machine learning algorithm can accurately predict the degree of causality in complex causation cases when the sample size of the dataset is large enough. We can predict the degree of causality of medical malpractice associated with other departments by re-selecting features and datasets.

Feature selection plays an essential role in model accuracy [38]. Regarding feature selection, we believe there are essential differences between different clinical departments in terms of medical factors, especially in technical issues [39]. At the same time, the different patient has different patient factors. Therefore, this study only included the data of orthopaedic-related medical malpractice cases and specialised and detailed the factors of doctors and patients as much as possible [25]: (1) Severe illness or progress rapidly; (2) Prognosis of disease; (3) Inadequate operation technique; (4) Damage consequence; (5) Inadequate internal fixation. The above five features are ranked in the top ten of the three models, which can inspire us. Orthopaedic surgeons need to pay more attention to technical issues, similar to the study by Liu et al. [40]. Of course, the patient's disease and development are also essential factors affecting responsibility identification.

At present, the determination of liability for medical malpractice relies on identification opinions, and the adoption rate of identification opinions by courts is very high [41], which is as high as 93.7%, according to the data in this paper. When issuing identification opinions, the experts' cognition of causality in the same case is biased due to their own background or experience [42, 43]. Ultimately, the identification opinions obtained by different experts in the same case may differ greatly, but it does not mean the different identification opinions are wrong. In order to standardise the identification opinions of medical malpractice evaluation, countries have drawn up a series of guidelines, such as "Guidance for judicial expertise of medical malpractice" (China SF/T 0097–2021) [24]; "The European Guidelines on medicolegal Methods of Ascertainment and Criteria of Evaluation" [44]. However, these guidelines do little to bridge the difference between identification opinions on the same case. Advanced computer technology has been widely used in forensic anthropology, dentistry and other disciplines [7, 45]. We expect to reduce cognitive bias through machine learning algorithms and models based on many Chinese case data. For example, in an orthopaedic medical malpractice case, after the expert has reached an identification opinion (causality), the case-related information (feature) is imported into the model, and the model predicts the most likely degree of causality (label) through a series of calculations. If the two are consistent, the expert identification opinion is more reliable.

The lack of medical and forensic knowledge puts judges in a difficult position when reviewing identification opinions [46]. The current solution is for experts to write their reports in plain language [47], or have the experts appear in court to explain their reports [48]. This study provides a new way to review identification opinions based on case data and computer algorithms. Machine learning can provide prediction and assessment of case outcomes, assist judges and lawyers to make more accurate decisions and reduce the influence of subjective factors. By reducing the interference of human factors, machine learning can improve the fairness and objectivity of court decisions and ensure the impartiality of justice. In particular, the data in this study comes from written judgement in a publicly available database. However, due to privacy issues and the fact that the court does not require the extraction of expert identification opinions in the judgment, and the effective information in the public database is limited, only 963 qualified cases are screened out of 13,245 judgments. More data for the courts, which have complete data on expert opinions, means that more accurate models can be built. Before this,

judges could not judge whether the expert opinions made by the commissioned experts were reasonable. Now, the machine learning prediction model formed by the modeling of a large amount of data can represent the mainstream opinion of most experts to a certain extent, which provides a basis for judges to compare. In this way, judges are no longer bogged down with a lot of medical knowledge and only need to compare the results of models with those of experts. If they are different, more scrutiny is needed.

The use of machine-learning algorithms in the justice system involves multiple ethical issues that require careful consideration in the deployment and use of these technologies [49, 50]. ML algorithms may inadvertently learn and amplify existing social biases, leading to unfair treatment of certain groups of people, such as those of a particular race, gender, or socioeconomic status [51]. So we did not include sex, race, and ethnicity in the characteristics, even though they might be important to the model. In addition, there is a growing demand to be able to "explain" ML systems' decisions and actions to human users, particularly when used in contexts where decisions have substantial implications for those affected and where there is a requirement for political accountability or legal compliance [52]. As mentioned earlier, we have adopted methods with high explanatory power for modeling. What's more, ML in the justice system need to process large amounts of sensitive data, including personally identifiable information, criminal records, and more. How to ensure the security and privacy of these data and prevent data leakage and misuse is an important ethical challenge. Besides, when machine-learning algorithms play a role in judicial decision making, how can responsibility be assigned if errors occur? Clear regulatory and legal frameworks for the use of machine learning in the justice system are lacking. This can lead to a lack of appropriate oversight and accountability mechanisms in practice. These issues remain to be resolved.

ML methodologies can be harnessed to categorize disability levels, employing a comparable approach to amass functional, disability, and health data pertinent to the ICF. This encompasses a spectrum of data, including personal attributes, medical documentation, and functional evaluations. The aggregated data undergo rigorous cleansing, transformation, and standardization to be utilized effectively in the training and prediction phases of machine learning algorithms. In alignment with ICF directives, pertinent features are meticulously selected and engineered to enable algorithms to accurately discern and prognosticate functional, disability, and health statuses. This endeavor may encompass sophisticated techniques such as natural language processing and image recognition. Despite the challenges posed by the scarcity of comprehensive databases, it is incontrovertible that in the burgeoning era of big data and artificial intelligence, forensic science stands poised for unprecedented advancements across its various disciplines.

5. Conclusion

This study used XGBoost, LightGBM and Random Forest for modelling. In the real data set, XGBoost performed best, and the model accuracy rate was 66%. In the virtual data set, the performance of XGBoost and LightGBM was the same, and the model accuracy rate was 80%. The overall accuracy of external verification was 72.7%. The optimal model was expected to predict the degree of causality accurately. The model established this time can only be used to predict the size of the causal relationship of orthopaedic-related medical injuries. We have verified the feasibility of this method and can further establish prediction models for other departments in future studies.

6. Limitation

In this study, the real data set is imbalanced, especially for the label classification of "Whole", and the insufficient data volume reduces the model's overall performance. The virtual data set

formed by oversampling may have problems with overfitting and data leakage. The sample size can be increased, and the model can be optimised by undersampling. Only three decision tree model-based machine learning integration algorithms are used, and others can be tried. In addition, we can try to combine automatic text classification technology with machine learning to reduce the workload [53, 54]. This paper does not explain the specific algorithm of the model. How does the computer form the optimal model through the algorithm?

Supporting information

S1 Table. Modeling data set and feature set.
(XLSX)

S1 File. Machine learning code and best models. They are available at https://github.com/nerdyqx/ML.
(ZIP)

S2 File.
(DOCX)

Author Contributions

Conceptualization: Qingxin Yang, Li Luo, Wenbo Zeng.

Data curation: Qingxin Yang, Zhangpeng Lin.

Formal analysis: Qingxin Yang.

Funding acquisition: Wei Wen.

Investigation: Qingxin Yang, Wei Wen.

Methodology: Qingxin Yang.

Software: Qingxin Yang.

Visualization: Qingxin Yang, Li Luo.

Writing – original draft: Qingxin Yang.

Writing – review & editing: Hong Deng.

References

1. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. PLoS Comput Biol. 2007; 3(6):e116. Epub 2007/07/03. https://doi.org/10.1371/journal.pcbi.0030116 PMID: 17604446; PubMed Central PMCID: PMC1904382.
2. Bianco SD, Parca L, Petrizzelli F, Biagini T, Giovannetti A, Liorni N, et al. APOGEE 2: multi-layer machine-learning model for the interpretable prediction of mitochondrial missense variants. Nature Communications. 2023; 14(1):5058. https://doi.org/10.1038/s41467-023-40797-7 PMID: 37598215
3. Dominguez Mantes A, Mas Montserrat D, Bustamante CD, Giró-i-Nieto X, Ioannidis AG. Neural ADMIXTURE for rapid genomic clustering. Nature Computational Science. 2023; 3(7):621–9. https://doi.org/10.1038/s43588-023-00482-7 PMID: 37600116
4. Bolton WJ, Wilson R, Gilchrist M, Georgiou P, Holmes A, Rawson TM. Personalising intravenous to oral antibiotic switch decision making through fair interpretable machine learning. Nature Communications. 2024; 15(1):506. https://doi.org/10.1038/s41467-024-44740-2 PMID: 38218885
5. Al-Zaiti SS, Martin-Gill C, Zègre-Hemsey JK, Bouzid Z, Faramand Z, Alrawashdeh MO, et al. Machine learning for ECG diagnosis and risk stratification of occlusion myocardial infarction. Nature Medicine. 2023; 29(7):1804–13. https://doi.org/10.1038/s41591-023-02396-3 PMID: 37386246

6. Lee Y-H, Won JH, Auh QS, Noh Y-K. Age group prediction with panoramic radiomorphometric parameters using machine learning algorithms. Scientific Reports. 2022; 12(1):11703. https://doi.org/10.1038/s41598-022-15691-9 PMID: 35810213

7. Galante N, Cotroneo R, Furci D, Lodetti G, Casali MB. Applications of artificial intelligence in forensic sciences: Current potential benefits, limitations and perspectives. Int J Legal Med. 2023; 137(2):445–58. Epub 2022/12/13. https://doi.org/10.1007/s00414-022-02928-5 PMID: 36507961.

8. Tack C. Artificial intelligence and machine learning | applications in musculoskeletal physiotherapy. Musculoskelet Sci Pract. 2019; 39:164–9. Epub 2018/12/07. https://doi.org/10.1016/j.msksp.2018.11.012 PMID: 30502096.

9. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. BMC medicine. 2015; 13:1. Epub 2015/01/08. https://doi.org/10.1186/s12916-014-0241-z PMID: 25563062; PubMed Central PMCID: PMC4284921.

10. Deo RC. Machine Learning in Medicine. Circulation. 2015; 132(20):1920–30. Epub 2015/11/18. https://doi.org/10.1161/CIRCULATIONAHA.115.001593 PMID: 26572668; PubMed Central PMCID: PMC5831252.

11. Bewes J, Low A, Morphett A, Pate FD, Henneberg M. Artificial intelligence for sex determination of skeletal remains: Application of a deep learning artificial neural network to human skulls. J Forensic Leg Med. 2019; 62:40–3. Epub 2019/01/15. https://doi.org/10.1016/j.jflm.2019.01.004 PMID: 30639854.

12. Li Y, Huang Z, Dong X, Liang W, Xue H, Zhang L, et al. Forensic age estimation for pelvic X-ray images using deep learning. Eur Radiol. 2019; 29(5):2322–9. Epub 2018/11/08. https://doi.org/10.1007/s00330-018-5791-6 PMID: 30402703.

13. Lai Y, Fan F, Wu Q, Ke W, Liao P, Deng Z, et al. LCANet: Learnable Connected Attention Network for Human Identification Using Dental Images. IEEE Transactions on Medical Imaging. 2021; 40(3):905–15. https://doi.org/10.1109/TMI.2020.3041452 PMID: 33259294

14. Scendoni R, Tomassini L, Cingolani M, Perali A, Pilati S, Fedeli P. Artificial Intelligence in Evaluation of Permanent Impairment: New Operational Frontiers. Healthcare. 2023; 11(14). https://doi.org/10.3390/healthcare11141979 PMID: 37510420

15. Vasudeva A, Sheikh NA, Sahu S. International Classification of Functioning, Disability, and Health augmented by telemedicine and artificial intelligence for assessment of functional disability. J Family Med Prim Care. 2021; 10(10):3535–9. Epub 2021/12/23. https://doi.org/10.4103/jfmpc.jfmpc_692_21 PMID: 34934642; PubMed Central PMCID: PMC8653435.

16. Li H, Dong S, Liao Z, Yao Y, Yuan S, Cui Y, et al. Retrospective analysis of medical malpractice claims in tertiary hospitals of China: the view from patient safety. BMJ Open. 2020; 10(9):e034681. Epub 2020/09/26. https://doi.org/10.1136/bmjopen-2019-034681 PMID: 32973050; PubMed Central PMCID: PMC7517568.

17. Knaak JP, Parzeller M. Court decisions on medical malpractice. Int J Legal Med. 2014; 128(6):1049–57. Epub 2014/03/29. https://doi.org/10.1007/s00414-014-0976-2 PMID: 24676889.

18. Liang F, Liu J, Zhou H, Liu P. Inequality in the last resort: how medical appraisal affects malpractice litigations in China. Int J Legal Med. 2021; 135(3):1047–54. Epub 2020/08/13. https://doi.org/10.1007/s00414-020-02386-x PMID: 32783158.

19. Zhang XD, Tian T, Yi XF, Sun JH. Comparison of Medical Dispute Resolution Mechanisms in China and Abroad. Fa Yi Xue Za Zhi. 2022; 38(2):150–7. Epub 2022/07/29. https://doi.org/10.12116/j.issn.1004-5619.2022.220106 PMID: 35899498.

20. Lacerte M, Forcier P. Medicolegal causal analysis. Physical Medicine and Rehabilitation Clinics of North America. 2002; 13(2):371–408. https://doi.org/10.1016/s1047-9651(01)00011-0 PMID: 12122852

21. Meilia PDI, Zeegers MP, Herkutanto, Freeman M. INFERENCE: An Evidence-Based Approach for Medicolegal Causal Analyses. Int J Environ Res Public Health. 2020; 17(22). Epub 2020/11/15. https://doi.org/10.3390/ijerph17228353 PMID: 33187384; PubMed Central PMCID: PMC7697841.

22. Meilia PDI, Freeman MD, Herkutanto, Zeegers MP. A review of causal inference in forensic medicine. Forensic Sci Med Pathol. 2020; 16(2):313–20. Epub 2020/03/12. https://doi.org/10.1007/s12024-020-00220-9 PMID: 32157581; PubMed Central PMCID: PMC7245596.

23. Liden M, Dror IE. Expert Reliability in Legal Proceedings: "Eeny, Meeny, Miny, Moe, With Which Expert Should We Go?". Science & justice: journal of the Forensic Science Society. 2021; 61(1):37–46. Epub 2020/12/29. https://doi.org/10.1016/j.scijus.2020.09.006 PMID: 33357826.

24. Cheng ZH, Zhang L, Wang L, Zhang J, Kong LJ, Yu L, et al. Comparation between Guidance for Judicial Expertise of Medical Malpractice and Medical Association Identification Rules of Medical Damage. Fa Yi Xue Za Zhi. 2022; 38(2):173–81. Epub 2022/07/29. https://doi.org/10.12116/j.issn.1004-5619.2022.220205 PMID: 35899501.

25. Thabet AM, Adams A, Jeon S, Pisquiy J, Gelhert R, DeCoster TA, et al. Malpractice lawsuits in orthopedic trauma surgery: a meta-analysis of the literature. OTA Int. 2022; 5(3):e199. Epub 2022/11/26. https://doi.org/10.1097/OI9.0000000000000199 PMID: 36425091; PubMed Central PMCID: PMC9580045.

26. Dong S, Shi C, Zeng W, Jia Z, Dong M, Xiao Y, et al. The Application of Graph Theoretical Analysis to Complex Networks in Medical Malpractice in China: Qualitative Study. JMIR Med Inform. 2022; 10(11): e35709. Epub 2022/11/04. https://doi.org/10.2196/35709 PMID: 36326815; PubMed Central PMCID: PMC9673000.

27. Yamamoto N, Sukegawa S, Watari T. Impact of System and Diagnostic Errors on Medical Litigation Outcomes: Machine Learning-Based Prediction Models. Healthcare (Basel). 2022; 10(5). Epub 2022/05/29. https://doi.org/10.3390/healthcare10050892 PMID: 35628029; PubMed Central PMCID: PMC9140545.

28. Heberer J, Eicher M. [Claims management from the perspective of the lawyer: Top 7 errors in medical liability law]. Unfallchirurg. 2020; 123(1):6–15. Epub 2019/11/07. https://doi.org/10.1007/s00113-019-00736-y PMID: 31690984.

29. Lv H, Li D, Li C, Yuwen P, Hou Z, Chen W, et al. Characteristics of the medical malpractice cases against orthopedists in China between 2016 and 2017. PLoS One. 2021; 16(5):e0248052. Epub 2021/05/13. https://doi.org/10.1371/journal.pone.0248052 PMID: 33979345; PubMed Central PMCID: PMC8115811.

30. Li H, Wu X, Sun T, Li L, Zhao X, Liu X, et al. Claims, liabilities, injures and compensation payments of medical malpractice litigation cases in China from 1998 to 2011. BMC Health Serv Res. 2014; 14:390. Epub 2014/09/15. https://doi.org/10.1186/1472-6963-14-390 PMID: 25218509; PubMed Central PMCID: PMC4261607.

31. Beyaz S, Acici K, Sumer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg. 2020; 31(2):175–83. Epub 2020/06/26. https://doi.org/10.5606/ehc.2020.72163 PMID: 32584712; PubMed Central PMCID: PMC7489171 to the authorship and/or publication of this article.

32. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022; 23(1):40–55. Epub 2021/09/15. https://doi.org/10.1038/s41580-021-00407-0 PMID: 34518686.

33. Du Y, Hua Z, Liu C, Lv R, Jia W, Su M. ATR-FTIR combined with machine learning for the fast non-targeted screening of new psychoactive substances. Forensic Sci Int. 2023; 349:111761. Epub 2023/06/17. https://doi.org/10.1016/j.forsciint.2023.111761 PMID: 37327724.

34. Brożek B, Furman M, Jakubiec M, Kucharzyk B. The black box problem revisited. Real and imaginary challenges for automated legal decision making. Artificial Intelligence and Law. 2023. https://doi.org/10.1007/s10506-023-09356-9

35. Hassija V, Chamola V, Mahapatra A, Singal A, Goel D, Huang K, et al. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation. 2024; 16(1):45–74. https://doi.org/10.1007/s12559-023-10179-8

36. Hassija V, Chamola V, Bajpai BC, Naren, Zeadally S. Security issues in implantable medical devices: Fact or fiction? Sustainable Cities and Society. 2021; 66:102552. https://doi.org/10.1016/j.scs.2020.102552.

37. Fisher A, Rudin C, Dominici F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J Mach Learn Res. 2019; 20. Epub 2019/01/01. https://doi.org/10.1080/01621459.1963.10500830 PMID: 34335110; PubMed Central PMCID: PMC8323609.

38. Chen X, He L, Shi K, Wu Y, Lin S, Fang Y. Interpretable Machine Learning for Fall Prediction Among Older Adults in China. Am J Prev Med. 2023; 65(4):579–86. Epub 2023/04/23. https://doi.org/10.1016/j.amepre.2023.04.006 PMID: 37087076.

39. Pröpper H. Viszeralmedizinische Schadensfälle: Analyse von 2763 viszeralmedizinischen Schadensfällen der Schlichtungsstelle für Arzthaftpflichtfragen der Norddeutschen Ärztekammern. Z Gastroenterol. 2014; 52(09):1050–61. https://doi.org/10.1055/s-0034-1366320 PMID: 25198084

40. Liu S, Zou J, Wang S, Liu G, Zhang Y, Geng S. Litigation analysis of medical damage after total knee arthroplasty: a case study based on Chinese legal database in the past ten years. Arthroplasty. 2022; 4(1):37. Epub 2022/10/01. https://doi.org/10.1186/s42836-022-00141-8 PMID: 36180903; PubMed Central PMCID: PMC9526297.

41. Loreto DBL, de Barros BAC, Rosa GCD, de Oliveira RN, Rosing CK, Fernandes MM. Analysis of Dental Case Reports in the Context of Court Decisions: Causal Nexus and Aspects of Fault. J Forensic Sci. 2019; 64(6):1693–7. Epub 2019/06/27. https://doi.org/10.1111/1556-4029.14089 PMID: 31237698.

42. MacLean CL. Cognitive bias in workplace investigation: Problems, perspectives and proposed solutions. Appl Ergon. 2022; 105:103860. Epub 2022/08/14. https://doi.org/10.1016/j.apergo.2022.103860 PMID: 35963213.

43. Thompson WC, Scurich N. How Cross-Examination on Subjectivity and Bias Affects Jurors' Evaluations of Forensic Science Evidence. J Forensic Sci. 2019; 64(5):1379–88. Epub 2019/02/23. https://doi.org/10.1111/1556-4029.14031 PMID: 30791101.

44. Ferrara SD, Baccino E, Bajanowski T, Boscolo-Berto R, Castellano M, De Angel R, et al. Malpractice and medical liability. European Guidelines on Methods of Ascertainment and Criteria of Evaluation. Int J Legal Med. 2013; 127(3):545–57. Epub 2013/04/09. https://doi.org/10.1007/s00414-013-0836-5 PMID: 23564275.

45. Thurzo A, Kosnacova HS, Kurilova V, Kosmel S, Benus R, Moravansky N, et al. Use of Advanced Artificial Intelligence in Forensic Medicine, Forensic Anthropology and Clinical Anatomy. Healthcare (Basel). 2021; 9(11). Epub 2021/11/28. https://doi.org/10.3390/healthcare9111545 PMID: 34828590; PubMed Central PMCID: PMC8619074.

46. Canela C, Buadze A, Dube A, Jackowski C, Pude I, Nellen R, et al. How Do Legal Experts Cope With Medical Reports and Forensic Evidence? The Experiences, Perceptions, and Narratives of Swiss Judges and Other Legal Experts. Front Psychiatry. 2019; 10:18. Epub 2019/03/01. https://doi.org/10.3389/fpsyt.2019.00018 PMID: 30814957; PubMed Central PMCID: PMC6381858.

47. Bali AS, Edmond G, Ballantyne KN, Kemp RI, Martire KA. Communicating forensic science opinion: An examination of expert reporting practices. Science & justice: journal of the Forensic Science Society. 2020; 60(3):216–24. Epub 2020/05/10. https://doi.org/10.1016/j.scijus.2019.12.005 PMID: 32381238.

48. Weng S. The dilemma and outlet of judicial determination of medical liability in China. Medicine, Science and the Law. 2023; 63(3):237–42. https://doi.org/10.1177/00258024231154816 PMID: 36748657

49. Papadouli V. Artificial Intelligence's Black Box: Posing New Ethical and Legal Challenges on Modern Societies. In: Kornilakis A, Nouskalis G, Pergantis V, Tzimas T, editors. Artificial Intelligence and Normative Challenges: International and Comparative Legal Perspectives. Cham: Springer International Publishing; 2023. p. 39–62.

50. Cingolani M, Scendoni R, Fedeli P, Cembrani F. Artificial intelligence and digital medicine for integrated home care services in Italy: Opportunities and limits. Front Public Health. 2022; 10:1095001. Epub 2023/01/24. https://doi.org/10.3389/fpubh.2022.1095001 PMID: 36684935; PubMed Central PMCID: PMC9849776.

51. Lo Piano S. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications. 2020; 7(1):9. https://doi.org/10.1057/s41599-020-0501-9

52. Coyle D, Weller A. "Explaining" machine learning reveals policy challenges. Science. 2020; 368 (6498):1433–4. https://doi.org/10.1126/science.aba9647 PMID: 32587011

53. Zhang B, Zhou S, Yang L, Lv J, Zhong M. Study on Multi-Label Classification of Medical Dispute Documents. Computers, Materials & Continua. 2020; 65(3):1975–86. https://doi.org/10.32604/cmc.2020.010914

54. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. J Forensic Leg Med. 2018; 57:41–50. Epub 2018/05/29. https://doi.org/10.1016/j.jflm.2017.07.001 PMID: 29801951.