scientific reports

Check for updates

OPEN

AI based predictive acceptability model for effective vaccine delivery in healthcare systems

Muhammad Shuaib Qureshi[1], Muhammad Bilal Qureshi[2], Urooj Iqrar[3], Ali Raza[4]✉, Yazeed Yasin Ghadi[5], Nisreen Innab[6], Masoud Alajmi[7] & Ayman Qahmash[8]✉

Vaccine acceptance is a crucial component of a viable immunization program in healthcare system, yet the disparities in new and existing vaccination adoption rates prevail across regions. Disparities in the rate of vaccine acceptance result in low immunization coverage and slow uptake of newly introduced vaccines. This research presents an innovative AI-driven predictive model, designed to accurately forecast vaccine acceptance within immunization programs, while providing high interpretability. Primarily, the contribution of this study is to classify vaccine acceptability into Low, Medium, Partial High, and High categories. Secondly, this study implements the Feature Importance method to make the model highly interpretable for healthcare providers. Thirdly, our findings highlight the impact of demographic and socio-demographic factors on vaccine acceptance, providing valuable insights for policymakers to improve immunization rates. A sample dataset containing 7150 data records with 31 demographic and socioeconomic attributes from PDHS (2017–2018) is used in this paper. Using the LightGBM algorithm, the proposed model constructed on the basis of different machine-learning procedures achieved 98% accuracy to accurately predict the acceptability of vaccines included in the immunization program. The association rules suggest that higher SES, region, parents' occupation, and mother's education have an association with vaccine acceptability.

Keywords Association Rule Mining, Childhood immunization, Feature importance, Machine learning, Vaccination, Vaccine Acceptance

To provide free routine vaccination to every child across the world, World Health Organization (WHO) started the immunization program in the year 1974 with an objective to reduce child mortality and morbidity rates. The program was also launched in Pakistan in the year 1978 aiming to protect newborns and mothers from nine lethal diseases through vaccination. The objectives of this program are (1) to make vaccines available for everyone, (2) to reduce vaccine preventable disease (VPD), and (3) to increase vaccine coverage throughout the country[1,2]. Initially, the vaccines included in this program were, Tetanus, Measles, Tuberculosis, Pertussis, Diphtheria, and Poliomyelitis. Later on, Hepatitis B, Hemophilic – Influenza type B, and Pneumococcal vaccines were also added in the years 2002, 2009, and 2012, respectively, while Inactivated Polio vaccine was added in the year 2015. Afterwards, COVID-19 and SARS-COV-2 vaccines were introduced to cover the COVID-19 pandemic where the vaccine acceptance rate was 90.6%. The report[3] shows that the life-warn threat changed the life behavior of the people, and also the economies around the earth. The main disagreement of the people towards vaccine acceptability was the infection possibility.

Unfortunately, Pakistan is still far behind the global immunization coverage target and has also missed out on the Millennium Development Goals (MDGs) for immunization coverage of 90% at the national level and 80% at district level[4]. According to Pakistan Demographic and Health Survey (PDHS 2017–18), the immunization coverage is still lower than expected and full immunization coverage at national level is just 66%

[1]School of Computing Sciences, Pak-Austria Fachhochschule Institute of Applied Sciences and Technology, Haripur, KPK, Pakistan. [2]Department of Computer Science & IT, University of Lakki Marwat, Lakki Marwat, KPK 28420, Pakistan. [3]Department of Computer Science, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Islamabad 46000, Pakistan. [4]Department of Computer Science, MY University, Islamabad, Pakistan. [5]Department of Computer Science, Al Ain University, 15551 Abu Dhabi, United Arab Emirates. [6]Department of Computer Science and Information Systems, College of Applied Sciences, AlMaarefa University, 13713 Diriyah, Riyadh, Saudi Arabia. [7]Department of Computer Engineering, College of Computers and Information Technology, Taif University, 21944 Taif, Saudi Arabia. [8]Department of Informatics and computer systems, College of Computer Science, King Khalid University, Abha, Saudi Arabia. ✉email: ali.raza.cs@myu.edu.pk; a.qahmash@kku.edu.sa

[5]. Figure 1presents province-wise full immunization coverage in Pakistan. Researchers from different fields have investigated the reasons for low immunization coverage and have found that vaccine acceptability is one of the key determinants of suboptimal immunization coverage[6,7]. Acceptability can be considered as one of the most important components of the sustainable immunization program. The individuals who refuse, delay or feel reluctant towards newly introduced vaccines, can be significant contributors of decreased immunization rates[8]. To increase the acceptability of vaccines included in immunization program, there is a need to investigate the factors associated with vaccine acceptability to provide immense insights for existing and newly introduced vaccines to ensure sustainability of ongoing and new vaccination programs. The factors can also be used as predictors of vaccine acceptability, including demographic characteristics, socioeconomic status, parental knowledge about vaccination and disease, attitude towards immunization, previous vaccination practices and access to information etc[9,10]. The most important factor in vaccine acceptability is parental awareness about the proper vaccination time and adverse effects on normal life in case of non-vaccination behavior[11]. The recent survey conducted by Sameer et al[12]., reported that the main barrier to vaccination is the lack of knowledge and fear about safety in the intake. In another study[13], 29–41% vaccine hesitancy was found in medical students regarding vaccine uptake fear. They have investigated possible associations between vaccine hesitancy and people.

The correlation among different factors like geographic variations, instability in people nature and impact of various patterns in patient's data are majorly due to the false positive rates in the data which impede the uptake of vaccine at right time[21]. In this paper, an action-oriented analytical approach is adopted, and interpretable vaccine acceptability prediction model is proposed. The proposed model categorizes vaccine acceptability as low, medium, partial high, and high on the bases of three main predictors of acceptance which are mother's attitude towards immunization, mother's previous vaccination practices and access to information. The association of several background characteristics (socioeconomic status (SES), parents' education, occupation, region, place of residence and quantity of children under the age of 5 in household) with predicted acceptability is discovered using association rule mining to provide valuable information to healthcare providers and policy makers for developing effective program strategies.

The key issues in the existing methods are as follows:

1. The existing literature lacks a predictive model capable of accurately assessing the acceptability of EPI vaccines. Furthermore, it does not thoroughly investigate the precise reasons behind the low acceptance of vaccines uptake.
2. Earlier studies have not clearly identified the factors that impact immunization coverage. Most research has pointed to vaccine acceptance as the primary determinant of low coverage. However, no highly interpretable technique has been proposed that can accurately predict vaccine acceptance based on a specific set of attributes.

The main contributions of the proposed study are as follows:

1. We proposed an interpretable vaccine acceptability prediction model using an action-oriented analytical approach. This model categorizes vaccine acceptability into four levels: low, medium, partial high, and high. The categorization is based on three primary predictors of acceptance: the expert attitude towards immunization, her previous vaccination practices, and her access to information, specifically her awareness of the impacts of vaccination. This approach also reveals the correlation between geographic variations and the instability in people's nature.
2. Our proposed model incorporates various strategies aimed at enhancing vaccine uptake and coverage, with careful consideration of regional differences and vaccine-specific factors. The model employs association rule mining to identify the relationships between background characteristics and vaccine defaulters.

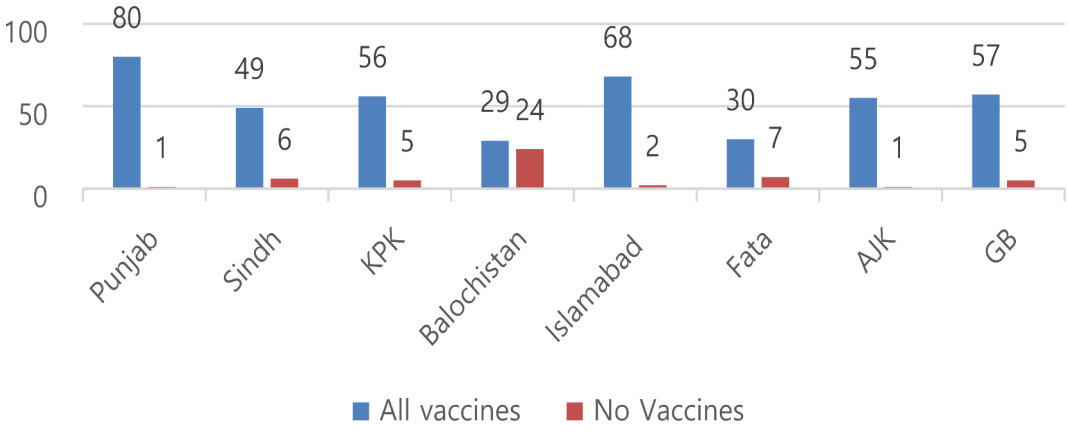

Fig. 1. Province-wise immunization coverage [3].

Related work

Before discussing existing works, we define terms in Table 1 used in the rest of the paper.

Earlier studies pointed out that there are several factors that influence vaccine uptake and results in low vaccination coverage. These factors are important to identify in order of increase immunization coverage. In a series of studies[1,2,22–25], descriptive data analytics is used so far in EPI program and other vaccination programs are discussed to analyze different aspects and factors that are playing an important role in impacting vaccine uptake. Such factors include the demographic, socio-demographic and non-socio-demographic characteristics. A number of reasons are discussed, and theories are proposed to help in understanding the impact of these factors on overall immunization coverage. This study presents the review of factors associated with vaccine acceptability such as demographic attributes, socio-economic attributes and other health related factors along with their impact on immunization coverage.

The earlier research focused on the impact of vaccine acceptance on coverage and identified it as the determinants of low coverage. Hagemann et al[26]., assessed the socio-demographic factors that impact parental decision-making regarding vaccination and stated that these factors could also differ region wise and vaccine wise. The study also revealed the factors associated with parental vaccine acceptance. Attendance at childcare units is also associated with higher acceptance and a higher level of parental education is negatively associated with vaccination. There is a need of different strategies to improve vaccine acceptance and coverage keeping in mind the difference of regions and vaccines. Mvula et al[27]., investigated that low vaccine acceptance/uptake has a strong association with low socioeconomic status, non-facility birth, distance from healthcare facilities and low maternal education. They analyzed data from a population-based birth cohort study to identify factors that are associated with the coverage and affect timely vaccine uptake for newly introduced pneumococcal and rotavirus vaccines. The study also examined the predictors of measles vaccine coverage and timeliness to inform national considerations of introducing second dose. The study also revealed the factors associated with vaccine uptake, infants born in farming families with lower literacy levels of mothers and those who live in remote areas are more likely to miss or skip the vaccines. Non facility birth, having more children in a family also affects the vaccine uptake. These factors are also considered to be strong predictors of vaccine uptake and acceptance. While introducing new vaccines, countries must ensure the adequate stock of vaccine, focus on the most vulnerable to improve access to immunization, ensure understandable and health appropriate information about vaccines is available in order to improve vaccine coverage.

There exists a plenty of literature on the factors associated with acceptance that can be used as a predictor of vaccine acceptance. The authors[7–10,28,29]evaluated knowledge about the disease, knowledge about vaccine, availability of vaccine, parents' behavior towards child vaccination and source of knowledge, parental practices towards mother and child vaccination and parents' demographics. Thus, the parental practices towards child vaccination and attitude towards immunization are important predictors of child immunization status. In all of the developed techniques for immunization coverage, the computer programming models different algorithms and data structures to diagnose and resolve different technical problems including text-to-code generation, technical query answering, prediction and optimization[30]. Similarly, IoT-based solutions lead towards potential mechanisms to address the challenges in most of the healthcare systems[31].

Handy LK, Maroudi et al[8]., explored the attitude, knowledge and believes regarding vaccines preventable diseases and vaccine acceptance among immunization providers and caregivers to provide insight about vaccine acceptability by investigating attitudes towards vaccines and the perceived influence of information source as

Acronym	Full Form
AI	Artificial Intelligence
PDHS	Pakistan Demographic and Health Survey
SES	Socioeconomic Status
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
EPI	Expanded Program on Immunization
WHO	World Health Organization
KPK	Khyber Pakhtunkhwa
MMR	Measles, Mumps, and Rubella
CV	Cross Validation
AUC	Area Under the Curve
FN	False Negative
FP	False Positive
TN	True Negative
TP	True Positive
XGBoost	Extreme Gradient Boosting
LightGBM	Light Gradient Boosting Machine
SVM	Support Vector Machine

Table 1. Acronyms and definitions.

new vaccines are introduced in immunization program. It also argues that vaccine acceptance is one of the important components of the sustainable immunization program, and knowledge about vaccine, vaccine safety and efficacy and disease can influence vaccine acceptance. So, there is a need to focus on these factors in order to understand which driver of acceptance can be leveraged to improve immunization program. The major part in vaccine uptake hesitancy is individual's belief in conspiracy theories which further can demonstrate distrust in science[32]. Another major factor of low vaccine acceptance is the antibody response which after first dose infects human body with other infection groups[33]. A survey conducted in America on about 6000 individual's shows that uncertainty in rural areas is majorly due to the untrusted beliefs in science and science-based policies[34]. In another study Teresa Morrone[7]evaluated the attitude and level of knowledge regarding meningococcal vaccine and investigated factors associated with vaccine acceptance. Parents characteristics (age, gender, marital status, educational level, employment status, no of the child's nationality), knowledge about the disease, knowledge about vaccine, availability of vaccine, parents' behavior towards child vaccination and source of knowledge are assessed. The analysis confirmed the positive impact of parent's higher education level, access to information, positive attitude towards vaccination and knowledge about vaccine are strong predictors of vaccine acceptance. The authors in[35]tried to investigate the underlying causes of vaccine acceptance, refusal and hesitation and concluded that primarily it begins with the parent-child relation. Shiferaw Birhanu et al[9]., accessed mother's knowledge, attitude and practices about child immunization and also explored the association of socio-demographic factors with immunization status and found that mother's education has a strong association with good attitude towards immunization and good practices have a strong association with information ever heard about immunization and its schedule, number of sessions needed and time when infant's immunization must be completed. To increase immunization coverage and to reduce child mortality rates, it is necessary to promote health education to increase parental knowledge regarding immunization and its practices. H. Harapan. H. et al[10]., assessed the knowledge, attitude, and practices regarding vaccine (KAP), history of dengue fever and other demographic factors to determine the role of vaccine acceptance explanatory variables. The factors associated with better vaccine acceptance are high monthly income, gender, high socio-economic-status, and good attitude towards dengue fever and vaccine practices.

Based on the above factors, different vaccine acceptance measurement tools and predictive models are also designed earlier. The Hoc group members have worked on designing regulatory framework that provides directions for standardizations in the field of AI in healthcare systems[36]. A systematic review conducted for investigating trust in crises was carried out where the quantitative synthesis showed that non-confidence in implementing vaccination plays a major role in people hesitation[37]. A. Bell et al[14]., proposed machine learning based early warning and monitoring system implemented as a dashboard to identify children at a great risk of not being vaccinated against MMR to facilitate healthcare workers and policy makers to take proactive decisions. A logistic regression model LASSO trained on child health records is used to identify vaccine hesitant families and then to predict vaccine uptake using child's vaccination records. The model is built as a web dashboard contains views at country, healthcare level and child level. It shows a risk in the form of scores using LASSO regression molding. The data from healthcare records consist of demographic and personal information features such as parental literacy level, age, marital status, work status, smoking status, number of children, and number of healthcare center visits, previous vaccine records and geographic information. EWS display results as an average risk score of students assigned to each healthcare center, individual risk score of each student and display child's values for features used for prediction. The level of risk of not getting MMR vaccine is divided into four levels which are low, medium, high, and very high. All the selected models are tested for accuracy, average log like hood, precision, and recall. The gradient booster tree model is better in terms of precision while LASSO regression is considered to be easier to interrupt because it used only 25 linearly combined features, thus it is considered ideal to build effective models for predicting vaccine hesitancy at the individual level. Niels Dalum Hansen et al[38]., demonstrated that prediction (vaccination uptake) of public health events can reduce reaction time of healthcare professionals and can improve vaccination uptake rates. They presented ensemble learning based model to predict vaccine uptake by combing the web and clinical data. An ensemble learning method called stacking is used to combine level-0 model into one prediction using mote model level-1. At first stage level-0 model is trained for final prediction and then using all the predictions of level-0 model as input. Level-1 model is experimenting with three different models: support vector regression (linear kernel), linear model and support vector machine (SVR) with Gaussian kernel. To predict with clinical data, three time series methods are used at level-0: AR models, Holt Winters & ARIMA while to predict with web data, linear model bagging and weighted majority. The actual vaccine uptake records from country official body and web queries related to vaccines are used for prediction. The combination of both yields lowest error, only web data gives error worse when used for prediction than for clinical data-based prediction.

The success of AI-based predictive modeling systems in healthcare for better judgment in certain functional areas is reported in[39]. The authors have concluded that the interpretations of the prediction outcomes are challenging. Shahmet al[16]., used predictive analytics to identify social and personal behaviors from large-scale data to predict future vaccine decision about influenza. First, it identified personal and social behavior patterns along with indicators of influenza vaccine uptake. The data include all demographic, socioeconomic attributes of the Israeli population. For each member demographic characteristics. Encounter with the Healthcare system, chronic illnesses, influenza vaccine history, and individual and family member's vaccination status, respiratory disease diagnoses, no of perceived medicines, hospitalization are compiled. Several machine learning methods are used for prediction and evaluated under the ROC curve. Results indicate that decisions related to vaccination can be explained as personal and social where personal dimension is shaped by specific behavior which can be affected by temporal factors. Out of several methods, XGBoost model achieved accuracy and a recall rate of 90% where ROC-AUC score is 0.91. Likewise, Sadaf Qazi et al[15]., also proposed a model for accurate prediction of defaulters from immunization program. Defaulters are those children who are at higher risk of missing their

next vaccine. A dataset from demographic and health survey (PDHS 2017-18) is used to carry out the study. The extracted features include demographic attributes (native language, region, and residence), socio-demographic attribute (SES), and immunization status of the child. Several machine learning methods are used to build the proposed model and experimentation is done to validate the model. Support vector machine, decision tree, naïve bayes, multilayer perceptron is used for model building. Conventional machine learning algorithms are favored for implementation in most of the healthcare systems including seasonal and non-seasonal diseases. Shaimaa et al., in[40]showed that different ML models achieve better accuracy rates in healthcare solutions, for example, SVM 93.4%, RF 93.3%, KNN 90.5%, and DT 87.9%. These higher accuracy rates reveal the effectiveness of implementing ML-based techniques. Certain performance measures (TP rate, TN rate, FP rate, FN rate accuracy, precision, recall, F-measure etc.) are used for each algorithm evaluation[41–47]. Multilayer perceptron is reported to achieve highest accuracy of 98% and 99.4% for AUC while the classification of defaulter as unvaccinated, partially low, partially medium, partially high and fully immunization. The proposed model also finds out the association of several background characteristics with defaulters using association rule mining. A number of 218 rules are generated using the apriori algorithm in SQL Server data tool to find out the association of demographic and sociodemographic attributes. Various machine learning algorithms are used for analyzing the discussion on social media and online blogs to highlight vaccine hesitant groups that refuse or delay the vaccine or prediction of future vaccine uptake[14,38]. These studies prove the feasibility to use machine learning models for immunization data to predict vaccine acceptability and also identify patterns and trends to help policy makers for designing effective strategies to improve immunization coverage[14–16,38]. Meanwhile, it also highlights the research gap as very little work is done in this domain and performed only a first step towards the understanding of power of predictive analytics for improvement of immunization coverage. There is also a need of highly interpretable model for accurate prediction of vaccine acceptance based on attributes set. Also, there is a need to find the association of demographic, socio-demographic and other factors with vaccine acceptance to understand the reasons of low vaccine acceptance that will help the policy makers to design effective strategies to improve immunization coverage rates.

Proposed framework for vaccine acceptability prediction

We have discussed various weaknesses in the available literature particularly a lack of predictive model that can predict the acceptability of vaccines included in EPI. In order to overcome the flaws, we have proposed a highly interpretable predictive framework that predicts vaccine acceptance and finds the association of acceptability with certain demographic and socio-demographic characteristics to understand the reasons behind low acceptance of vaccines. The proposed framework is shown in Fig. 2 for accurate prediction of vaccines acceptability and association of acceptability with background characteristics.

Methods

Data set description

Demographic and vaccination data extracted from PDHS (2017-18)[5] is used in this paper. The demographic & health survey is conducted to provide data for monitoring and impact evaluation indicators in the area of population, maternal and child health issues, and nutrition in Pakistan. The dataset contains 50,486 data records and 1186 attributes.

Data preprocessing

The demographic and health data is inconsistent or error prone. Particularly, it is observed that the child's vaccination records have significant missing data. For some respondents, entire demographic and vaccination records are missing. This inconsistent and inaccurate data put negative impact on results. Preprocessing of the inconsistent data is carried out and the data is first cleansed from outliers and irrelevant values. The instances with missing values in child's immunization and mother's vaccination records are removed and other demographic attributes with missing values are handled using simple imputation method called mean/max method. Missing values are filled with variable mean values for continuous variables. Finally, the continuous valued attributes such as maternal age is categorized into three groups: young, middle, and old. To further clarify, the maternal age 15–26 is categorized as "young", 27–36 as "middle" and 37–49 as "old".

Data cleaning

The dataset contains incomplete values such as Null value, blank space, or some special character "?" which needs to be replaced or removed to improve results. Immunization records in the dataset can only have two possible values: either child is vaccinated against disease or not. The missing values in immunization and health related records cannot be replaced but can only be handled by ignoring and dropping the instance. The rows with missing immunization and health data are removed from the dataset while missing records in demographic attributes are handled using the mean/mode method.

Data discretization

Discrete attributes increase information representation and minimize memory usage. This makes the data simpler to understand and knowledge retrieval methods become faster. The data set contains numerous continuous-valued attributes which need to be discretized into subsets of categorical values. One of the continuous-valued attributes in this hypothetical dataset is 'Maternal age' which contains values from 15 to 49. The values are first divided into three sets of ages i.e., 15–26, 27–36 and 37–49 and then named as young, middle and old.

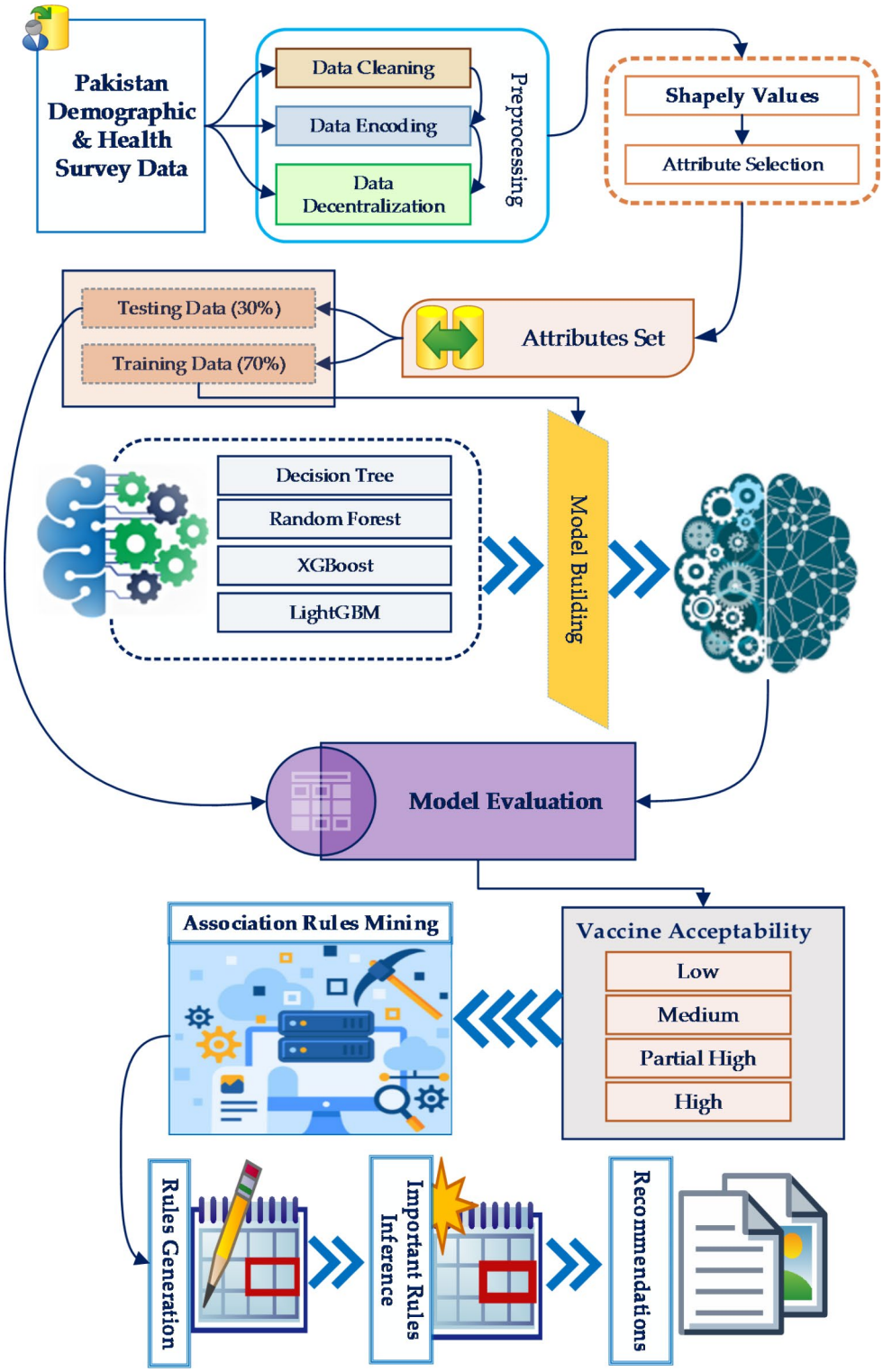

Fig. 2. AI-based framework for predicting vaccine acceptability.

Attribute selection

The PDHS dataset[5] comprises of 1186 attributes, but not all the attributes were required for experimentation. Thus, only the relevant attributes are extracted using domain knowledge. After preprocessing the dataset consist of 31 attributes and 7151 data instances. Table 2 shows the extracted demographics, socio-demographic and health attributes along with each attribute's value.

Sr. No	Attribute Name	Attribute Values
1	Region	1 = Punjab, 2 = Sindh, 3 = KPK, 4 = Baluchistan, 5 = Fata, 6 = AJK, 7 = Gb
2	Residence	1 = Urban, 2 = Rural
3	Maternal age	1 = Young, (15–26) 2 = Middle(27–36), 3 = Old (37–49)
4	Marital status	1 = Married, 2 = Widow, 3 = Divorced, 4 = No longer living together/separated
5	Parental education	0 = No education 1 = Primary, 2 = Secondary, 3 = Higher
6	Occupation	0= Did not work, 1 = Professional/Technical/Managerial, 2 = Clerical, 3 = Sales, 4 = Agricultural - Self-employed, 7 = Services, 8 = Skilled Manual, 9 = Unskilled Manual
7	No of children under 5	-
8	Place of Child's birth	10 = Home, 20 = Public Sector, 30 = Private Sector, 96 = Other
9	Socio-Economic Status	1 = Poorest, 2 = Poorer, 3 = Middle, 4 = Richest, 5 = Richer
10	Frequency of watching TV	0 = No, 1 = Yes
11	Frequency of reading newspaper	0 = No, 1 = Yes
12	Frequency of using internet	0 = No, 1 = Yes
13	Frequency of listening radio	0 = No, 1 = Yes
14	Own a mobile	0 = No, 1 = Yes
15	Tetanus injection during pregnancy	0 = No, 1 = Yes
16	Antenatal visits during pregnancy	0 = No, 1 = Yes
17	Baby's Postnatal checkups	0 = No, 1 = Yes
18	Vaccination through EPI card	0 = No, 1 = Yes
19	Vaccination (BCG, Polio…)	0 = No, 1 = Yes

Table 2. Extracted attributes values.

Attributes	Attribute Values	New Mapped Attribute	Attribute Values
Infant immunization practice by EPI card	1 = Yes. 0 = No	Previous Practices	0 = bad 1 = Average 2 = Good
Vaccination practices during pregnancy	1 = Yes, 0 = No		
Use of antenatal care	1 = Yes, 0 = No		
Postnatal Checkup	1 = Yes, 0 = No		
Frequency of watching television	1 = Yes, 0 = No	Access to Information	0 = No 1 = Yes
Frequency of reading newspaper	1 = Yes, 0 = No		
Frequency of listing radio	1 = Yes, 0 = No		
Has mobile phone	1 = Yes, 0 = No		
Frequency of using Internet	1 = Yes, 0 = No		
BCG	1 = Yes, 0 = No	Attitude towards Immunization	0 = Poor 1 = Bad 2 = Average 3 = Good 4 = Very Good
Polio 0, Polio 1, Polio 2, Polio 3	1 = Yes, 0 = No		
Pentavalent 1, Pentavalent 2 Pentavalent 3	1 = Yes, 0 = No		
Measles 1, Measles 2	1 = Yes, 0 = No		
Pneumococcal 1, 2, 3	1 = Yes, 0 = No		

Table 3. Attributes mapping.

Attribute mapping

The predictive features of vaccine acceptability are mapped into three new attributes: access to information, attitude towards immunization and previous practices. All these new attributes are formed by mapping those attributes which have a potential to measure these attributes. This mapping is done on the bases of WHO guidelines and previous studies[1,7–9,38]. The criteria of attributes mapping are depicted in Table 3.

Model building

Handling imbalanced dataset

In imbalance classification, the class distribution is not uniform among classes and a number of instances per class is not evenly distributed in the target variable. In predictive modeling, imbalance classification poses a threat as most of classification algorithms are designed on the assumption of an equal number of examples for each class. Thus, imbalance distribution can result in poor predictive performance.

This data imbalance problem is handled using Synthetic Minority Over-sampling Technique (SMOTE)[48]. SMOTE performs efficiently in addressing class imbalance issues in the datasets. In the case of predictive modeling, the class imbalance issues occur when the number of instances across different classes is not distributed evenly which results in poor predictive performance for the minority class. SMOTE generates synthetic samples

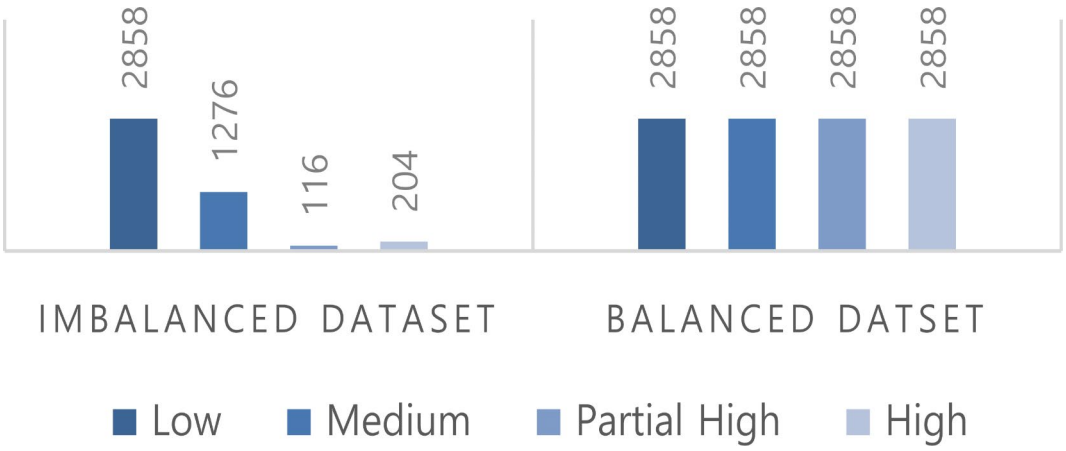

Fig. 3. Class distribution of instances before and after SMOTE.

Imbalanced Dataset					Balanced Dataset				
Models	**Accuracy (%)**	**Precision(%)**	**Recall(%)**	**F1-Score(%)**	**Models**	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**F1-Score(%)**
Decision Tree	99.5	100	100	100	Decision Tree	96	97	96	97
Random Forest	99.9	100	100	100	Random Forest	97	98	88	98
XGBoost	99.8	100	100	100	XGBoost	97	98	98	98
LightGBM	99.9	100	100	100	LightGBM	98	98	98	98

Table 4. Performance metrics.

for the minority class rather than to duplicate the existing ones. The duplicated observations do not add any information to the model and can be synthesized from existing observations. The synthesis can be done by selecting two or more similar instances and generating new instances that are combinations of these instances, hence helps in balancing the class distribution. This allows the predictive model to learn equally from all classes and thus improving the overall performance of the predictive model. Figure 3 depicts the class distribution of dataset before and after synthetic minority over-sampling.

Modeling and evaluation

In this research work, we put greater emphasis on the high interpretability and performance of the prediction matrix[49]. A number of machine learning algorithms (Decision Tree, Random Forest[50,51], XGBoost[42], and LightGBM[52]) are applied in Jupyter Lab on the dataset for prediction model building. Jupyter Lab is a web-based interactive development environment for Jupyter Notebooks, code, and data. The actual language of Jupyter Lab is Python; however, it supports many languages through its use of different kernels like R, Julia, and others. The interface and extensions are also developed using JavaScript.

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. When it comes to computation, LightGBM is considered to be a very powerful and fast processing algorithm. The algorithm is called light because of its speed and computation power. It has the capability to deal with large amount of data and it also takes less memory to run. One of the main reasons to use this algorithm is its faster training speed and higher efficiency. Also, it has better accuracy than any other boosting algorithm. It produces complex tress by following leaf wise split approach which is the main factor in achieving higher accuracy.

For model building, the data of KPK region hold out for model validation, rest of 85% data are further divided into training and testing dataset[53]. In this model, we evaluated the predictive results using 10-fold Cross-validation (CV) test. The mean value of the 10-fold CV test was obtained by repeating the Stratified loop procedure 100 times. Where the data are randomly selected, using 7 folds for training purposes and remaining 3 folds are used for testing. The performance of each model is measured with different performance measures. Each model is also validated on the 15% of regional dataset using same performance measures. Tables 4 and 5present the outcomes of different performance measures for applied algorithms on both datasets. The performance measures used for model evaluation are accuracy, precision, recall and F1-measures[54].

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	98.6	99	99	99
Random Forest	93	94	93	93
XGBoost	98.6	99	99	99
LightGBM	99	100	99	99

Table 5. Performance metrics for each algorithm on KPK data.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

where.

TP (True Positive) refers to the number of correctly predicted positive cases. In the context of vaccine acceptability, this would mean the number of instances where the model correctly predicted high or acceptable vaccine coverage when it was indeed the case.

TN (True Negative) is the number of correctly predicted negative cases. This refers to instances where the model correctly predicted low vaccine acceptability when that was the true scenario.

FP (False Positive) is the number of incorrectly predicted positive cases. In this case, it refers to instances where the model predicted high vaccine acceptability, but in reality, the acceptability was low.

FN (False Negative) is the number of incorrectly predicted negative cases. It refers to the cases where the model predicted low vaccine acceptability, but the true value was high.

Feature importance

The aim of this research is to develop a highly interpretable model that is highly accurate in prediction of vaccine acceptability. The selection of attributes set is performed using domain knowledge and feature importance test using SHAP[55,56], which is very fast method to estimate each feature's importance for the model. First, SHAP values are calculated using SHAP Library and results are plotted using SHAP-Tree Explainer. These decision plots show how a model arrives at their predictions and make the model highly interpretable. The pseudo code of shapely value algorithm is presented in Algorithm 1.

Figure 4 represents the most significant attributes based on the calculation of SHAP values for each attribute presented in descending order. Similarly, Fig. 5 shows the most important features based on the computation of the SHAP values for each feature on sample of total 1072 test points. Points are colored according to the value of the feature for each class. The individual SHAP values of each attribute for each predictive class are presented in Table 6. Individual SHAP value force plot is used to represent SHAP values for each of attribute at any single instance. The individual SHAP value force plot is created for few randomly selected instances to describe the impact of individual instance values[57]. It represents those features that contribute to push the prediction of base value. The SHAP value force plot shows how each feature in the dataset equally contributes to pushing the prediction from the base value to the model's output for individual instances. The base value 0.03 is the average model output over the test dataset that we passed. The procedure for creating SHSP value force plot is that first the SHAP values are generated for each of the feature in the predictive model using the SHAP built-in library. This calculation involves determining the contribution of each feature to the prediction. The base value that is 0.03 in our case is the average model output over the training dataset. The individual SHAP value force plot represents the SHAP values for each attribute at a single instance. The features that push the prediction higher; are represented in red color and features that push the prediction lower; are represented in blue color. This colorful representation helps to understand the impact of each feature on the model's prediction for a specific instance. Access to information push prediction towards higher in LightGBM as presented in Fig. 6.

Two of the decision plots used to plot the results are variable importance plot (For global interpretability) and individual SHAP value force plot (For local interpretability).

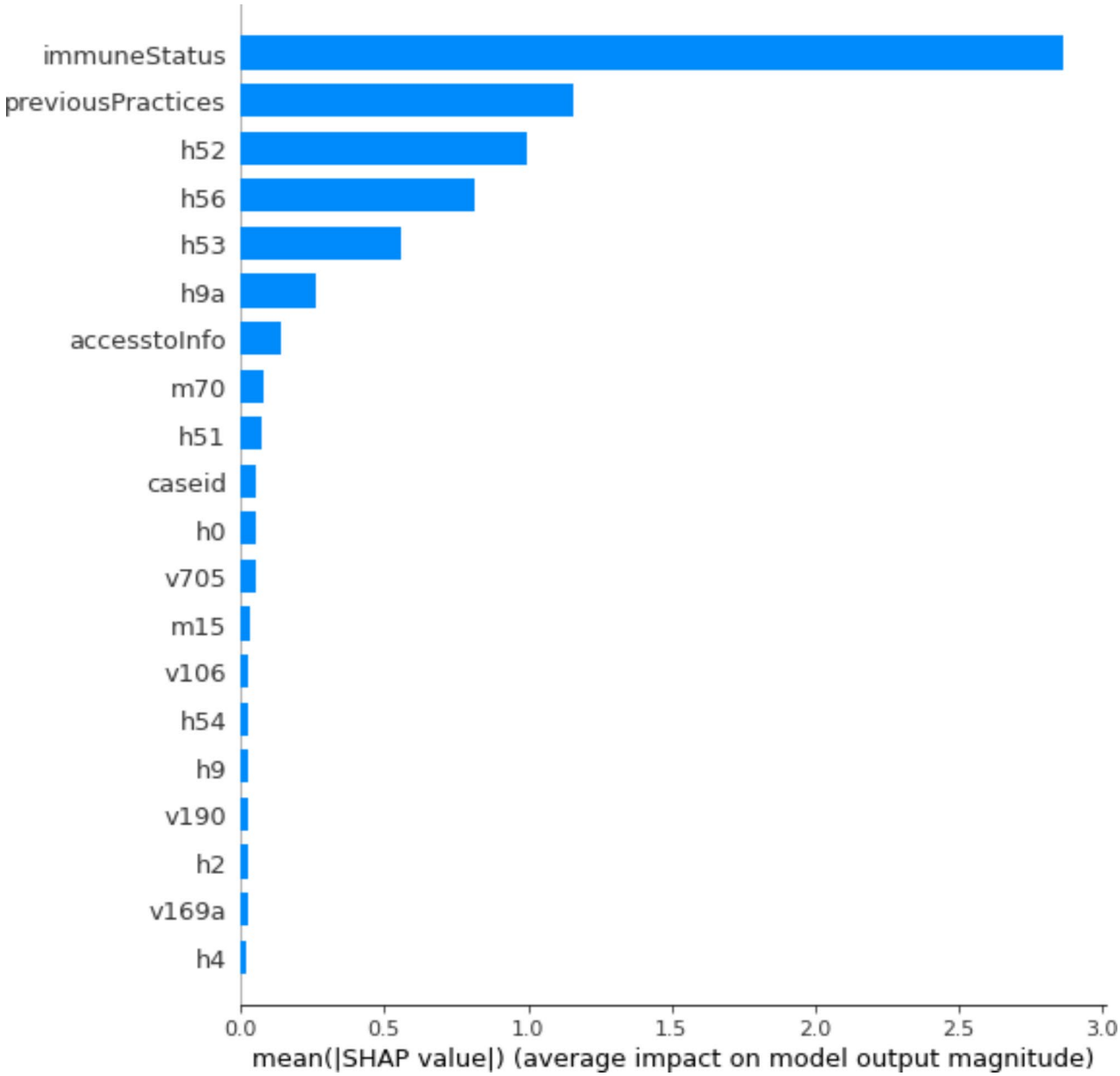

Fig. 4. Local interpretation-LightGBM.

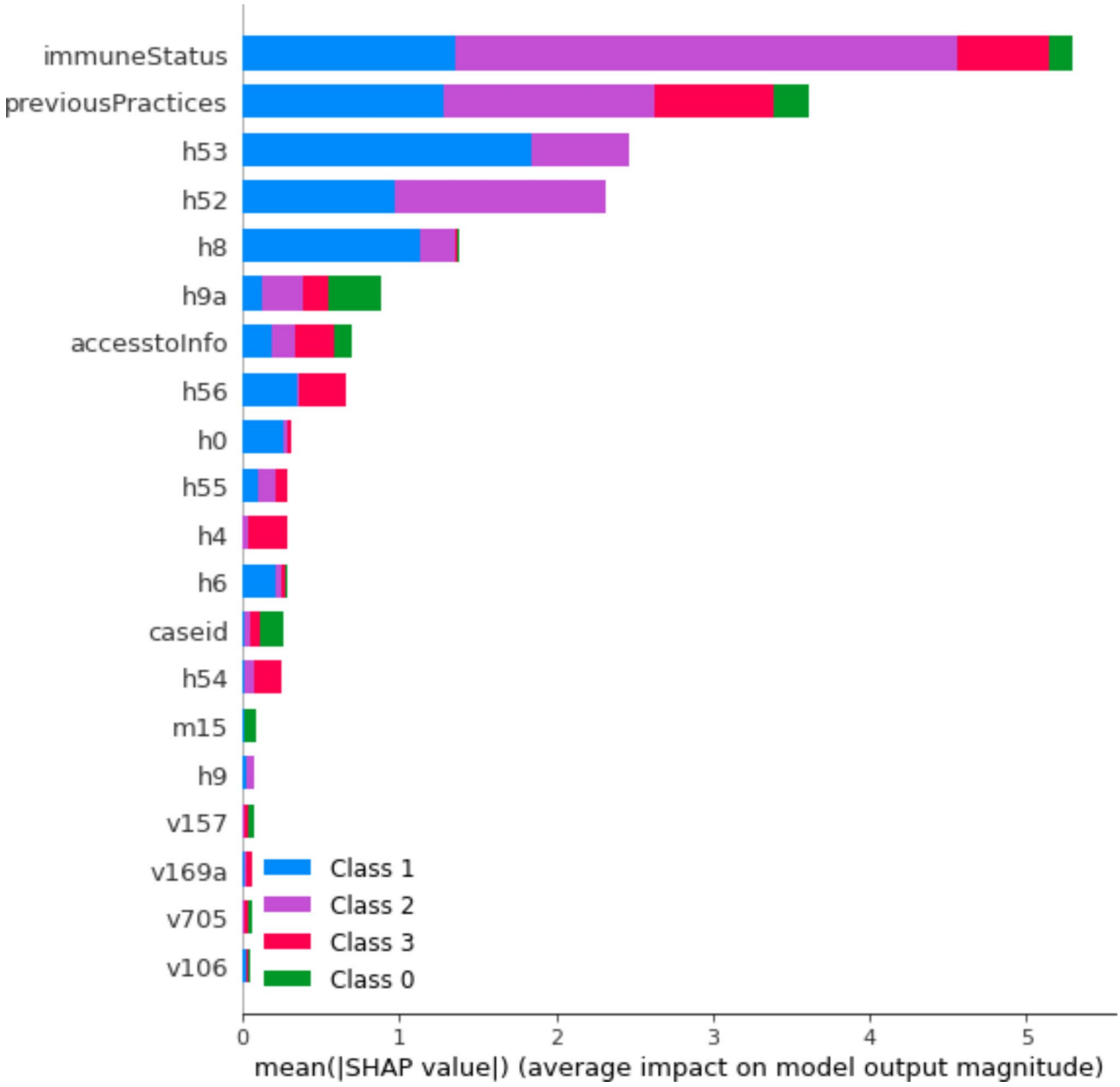

Fig. 5. Global interpretation-LightGBM. H9a (Has health card), H56 (Pneumococcal-3), H6 (Polio-2), H4 (Polio-1), V024 (Region), V137 (no of children under 5), H54 (Pneumococcal-1), H8 (Poio-3).

Class-Low		Class- Medium	
Features	**Weight**	**Features**	**Weight**
H53	1.8	Attitude towards immunization	2.86
Attitude towards immunization	1.63	Previous practices	1.14
Previous practices	1.3	H52	0.99
H56	1.15	H53	0.55
H8	0.66	H56	0.82
H55	0.31	H9a	0.259
H52	0.30	Access to information	0.13
Access to information	0.18		
H54	0.14		
H0	0.09		
Class-High		**Class-Partial High**	
Features	**Weight**	**Features**	**Weight**
H9a	0.40	Previous practices	0.74
Previous practices	0.26	Attitude towards immunization	0.48
Attitude towards immunization	0.21	H8	0.38
Access to information	0.11	H9a	0.34
V137	0.07	Access to information	0.26
V024	0.06	H54	0.14
H56	0.01	H0	0.13
H6	0.01	H56	0.06
		H4	0.05

Table 6. SHAP values. H9a (Has health card), H56 (Pneumococcal-3), H6 (Polio-2), H4 (Polio-1), V024 (Region), V137 (no of children under 5), H54 (Pneumococcal-1), H8 (Poio-3), H0 (Polio-0).

Fig. 6. Local interpretation – LightGBM.

Input: Training Dataset

Output: Importance Value for each feature

1. Initialize all the weight values Ji to 0;
2. ***foreach*** attribute ***do***
3. Create a power set from all coalitions. $\{\pi_1, \ldots\ldots, \pi_k\}$ over F;
4. ***foreach*** $\pi j \in \{\pi_1, \ldots\ldots, \pi_k\}$ ***do***
5. Calculate Δi;
6. **end**
7. Calculate Øi;
8. **end**

Algorithm 1. Pseudocode of shapely value algorithm.

Sr. No.	Rules	Class	Lift	Confidence interval (%)
1	Region = Baluchistan, SES = Poorer, Maternal Education = No Education	Low	1.4	94
2	Region = Fata, Maternal education = No Education, Marital Status = Married	Low	1.38	93
3	Region = Sindh, SES = Poorer, No of children under 5 = 2, Maternal Education = No Education, Marital Status = Married	Low	1.3	87
4	SES = Poorer, Maternal Education = Higher, Place of child birth = Private Hosp	Low	1.82	89
5	Occupation = unskilled manual, Father's education = Middle, Marital status = Married, No of children under 5 = 3, SES = poor	Low	1.86	86
6	SES = poor, Residence = Rural, Marital Status = Married, Maternal age = Middle, Education = Middle	Low	1.49	85
7	Maternal age = Young, Marital Status = Married, Education = Secondary	Low	1.55	97.5
8	Marital Status = Married, Residence = Urban, Education = Secondary	Medium	1.96	91.23
9	Type of Residence = Urban, Education = Secondary, Place of child birth = Govt Hosp, Marital status = Married	Medium	1.35	90.34
10	Mother's age = Middle, Marital Status = Married, Education = Secondary, no of children = 2	Medium	2.52	91.50
11	Region = ICT, Residence = urban, Marital status = Married, Occupation = professional	Medium	1.02	71
12	Region = Punjab, Mother's education = No education, Mother's age Middle	Medium	1.96	78

Table 7. Association rules.

Results and discussion

The immunization coverage of KPK reported in PDHS is only 55%. According to the target set by *the World Health Organization*, Immunization coverage must be 80% at district level. So, we divided immunization coverage rates according to WHO guidelines as more than 80% acceptability is considered to be High, acceptability ranges from 70 to 80% is considered to be Partial High, acceptability below than 60% is considered as Low and acceptability ranges from 60 to 75% is considered as Medium. However, there could be other factors that could affect immunization coverage rates.

Association rule mining

The Association rules are generated using SPSS (Statistical Package for Social Sciences). We applied apriori algorithm to generate association rules. Apriori algorithm is well-known exhaustive algorithm. The association of demographic and socio-demographic attributes is analyzed with low and medium vaccine acceptability. Using SPSS, total 996 rules are generated based on minimum support value of 3. The sample of rules generated from *Pakistan Demographic & Health Survey dataset* is depicted in Table 7. The association of attributes can be easily interpreted from the rules reported in the following: rules 1, 3, and 4 suggest that the low vaccine acceptability has direct association with socioeconomic status (SES). If SES is poorer, the family would be less likely to immunize their child. So, if the wealth index of the family improves then there are chances of vaccination acceptability for newly added and existing vaccines will also improve. Another attribute "maternal education i.e., maternal literacy level" is reported to have a significant association with knowledge regarding infant immunization and a positive attitude towards immunization[1,7,9,11,58]. However, *Hage Mann et al.*[26], reported negative association of parental education with vaccine acceptance. This research focused both mother's and father's literacy level for analysis and also confirms the fact that if parenteral literacy level is primary or low, there will be higher chances of low vaccine acceptability. It makes the fact evident that parental education has a positive association with parental knowledge and their attitude towards immunization. Hence, the lowest literacy level negatively affects parental attitude and knowledge towards immunization that results in lower vaccine acceptance.

Earlier studies also identified that there are considerable variations in immunization coverage across the regions in Pakistan as there persist large inequalities in vaccination uptake[1,4,5,11,26,58]. In this research, the generated association rules also confirm that region has a correlation with vaccine acceptability and if a region is a remote or undeveloped area of Baluchistan, Fata or Sindh, there will be low vaccine acceptability. Additionally, parent's occupation can also be considered as an important factor that contributes towards decision to vaccinate a child. As it is explained earlier that parent's occupation is an important socio-demographic factor that influences vaccine acceptability at large. The vaccination decision can be affected by parent's occupation in several ways. For example, parents with high socio-economic status have easy access to different healthcare services that can lead to higher vaccination rates. Similarly, parents in professional or managerial roles have better exposure to the health information and resources. The healthcare work environment that involves interactions with the healthcare workers have increased awareness about the importance of vaccinations. Strangely, no interesting associations of attributes: "age of mother" and "number of children under five in household" has been found in this study.

The findings from this study provide valuable insights into the demographic and socioeconomic factors influencing vaccine acceptability. These insights can inform policymakers in designing tailored interventions for populations that are less likely to get immunized, such as poorer households or those in rural areas. By leveraging predictive analysis, healthcare programs can adopt a data-driven approach to identify high-risk groups, enabling targeted outreach efforts, such as mobile vaccination clinics, education campaigns, and financial support programs to reduce barriers to vaccine access. Furthermore, policy adjustments can be made to ensure equitable access to vaccines, ensuring that resource allocation is based on the specific needs of the population, thereby improving overall immunization coverage and contributing to public health goals.

Conclusion and future work

An interpretable vaccine acceptability prediction model is proposed in this paper, and a number of classifiers like Random Forest, XGBoost, Decision Tree and LightGBM are applied on the PDHS dataset. The outcomes indicates that the proposed model achieved 98% accuracy for predicting acceptability as low, medium, partial high and high. The proposed framework is a step towards a data-driven approach that provides a set of machine learning techniques to use predictive analytics. Hence, predictive analysis can strengthen vaccination programs by accelerating targeted measures to increase vaccine acceptance. Predictive analysis provides insights into which demographic, socioeconomic, and behavioural factors are most strongly associated with vaccine hesitancy or low vaccine acceptability. By identifying these patterns, policymakers and healthcare providers can design tailored interventions that focus on specific subsets of the population most likely to refuse or delay vaccinations, such as low-income families or those in rural areas. It also helps in predicting which regions or groups are at higher risk of low vaccine uptake, resources can be more efficiently allocated, ensuring that immunization campaigns target the right communities with appropriate messaging, outreach, and healthcare access. Additional major contributions of this research are high interpretability of model and association rule mining of acceptability with their contextual features which emphasized the effect of demographics and socioeconomic attributes on vaccination acceptability. Predictive analytics and exploration of association rules can play an important role with its limited application in health care, particularly in EPI. To add other behavioral aspects of parents such as knowledge of vaccination and childhood diseases, use of bigger real-world datasets and validation of model for all region's data separately can be considered as additional future extensions of the proposed work.

Limitations of the presented research study

While this study provides valuable insights into vaccine acceptability using predictive analysis, several limitations are following:

Generalization to other regions

The findings from this study are based on data from Pakistan, specifically from the PDHS dataset. Therefore, the results and the predictive model may not generalize well to other countries or regions with different healthcare systems, vaccine policies, and demographic factors. Future studies should test this model with datasets from various regions to evaluate its broader applicability.

Interpretability of the model

While we employed SHAP to enhance the interpretability of the model, machine learning models, particularly complex algorithms like LightGBM, can still be difficult for non-technical stakeholders to fully understand. Policymakers and healthcare providers may require additional tools or simplified models to facilitate actionable insights.

Data availability

The data used in the study are publicly available via a github link https://github.com/Intelligent-models/AI-Pred-Vaccine.

Received: 9 June 2024; Accepted: 17 October 2024
Published online: 04 November 2024

References

1. Imran, H. et al. Routine immunization in Pakistan: comparison of multiple data sources and identification of factors associated with vaccination. *Int. Health.* **10**, 84–91 (2018).
2. Noh, J. W. et al. Factors affecting complete and timely childhood immunization coverage in Sindh, Pakistan; a secondary analysis of cross-sectional survey data. *PloS One.* **13**, e0206766 (2018).
3. Graffigna, G., Palamenghi, L., Barello, S. & Stefania, B. Cultivating acceptance of a COVID-19 vaccination program: lessons from Italy. *Vaccine.* **38**, 7585 (2020).
4. Butt, M., Mohammed, R., Butt, E., Butt, S. & Xiang, J. Why have immunization efforts in Pakistan failed to achieve global standards of vaccination uptake and infectious disease control? *Risk Manage. Healthc. Policy.* **13**, 111–124 (2020).
5. Demographic, C. Health Survey, Key Findings. *Phnom Penh and Calverton, Maryland, USA: Nationanl Insitute of Statistics* (2011).
6. Sharma, S., Akhtar, F., Singh, R. K. & Mehra, S. Understanding the three as (awareness, Access, and acceptability) dimensions of vaccine hesitancy in Odisha, India. *Clin. Epidemiol. Global Health.* **8**, 399–403 (2020).
7. Morrone, T., Napolitano, F., Albano, L. & Di Giuseppe, G. Meningococcal serogroup B vaccine: knowledge and acceptability among parents in Italy. *Hum. Vaccines Immunother.* **13**, 1921–1927 (2017).
8. Handy, L. K. et al. The impact of access to immunization information on vaccine acceptance in three countries. *PloS One.* **12**, e0180759 (2017).
9. Birhanu, S., Anteneh, A., Kibie, Y. & Jejaw, A. Knowledge, attitude and practice of mothers towards immunization of infants in health centres at Addis Ababa, Ethiopia. *Am. J. Health Res.* **4**, 6–17 (2016).
10. Harapan, H., Anwar, S., Setiawan, A. M. & Sasmono, R. T. Dengue vaccine acceptance and associated factors in Indonesia: a community-based cross-sectional survey in Aceh. *Vaccine.* **34**, 3670–3675 (2016).
11. Crouch, E. & Dickes, L. A. A prediction model of childhood immunization rates. *Appl. Health. Econ. Health. Policy.* **13**, 243–251 (2015).
12. Gopalani, S. V. et al. Barriers and factors associated with HPV vaccination among American indians and Alaska Natives: a systematic review. *J. Community Health.* **47**, 563–575 (2022).
13. Lo Moro, G., Cugudda, E., Bert, F., Raco, I. & Siliquini, R. Vaccine hesitancy and fear of COVID-19 among Italian medical students: a cross-sectional study. *J. Community Health.* **47**, 475–483 (2022).
14. Bell, A. et al. in. *IEEE International Conference on Healthcare Informatics (ICHI).* 1–6 (IEEE). (2019).
15. Qazi, S., Usman, M. & Mahmood, A. A data-driven framework for introducing predictive analytics into expanded program on immunization in Pakistan. *Wien. Klin. Wochenschr.* **133**, 695–702 (2021).

16. Shaham, A., Chodick, G., Shalev, V. & Yamin, D. Personal and social patterns predict influenza vaccination decision. *BMC Public. Health.* **20**, 1–12 (2020).
17. Omotunde, H. & Mouhamed, M. R. The Modern Impact of Artificial Intelligence Systems in Healthcare: A Concise Analysis. *Mesop. J. Artif. Intell. Healthc.* 66–70 (2023).
18. Karne, R. & Sreeja, T. Clustering algorithms and comparisons in vehicular ad hoc networks. *Mesopotamian J. Comput. Sci.* **2023**, 115–123 (2023).
19. Zhou, Y. et al. Dermatophagoides pteronyssinus allergen Der p 22: cloning, expression, IgE-binding in asthmatic children, and immunogenicity. *Pediatr. Allergy Immunol.* **33**, e13835 (2022).
20. Cao, P. & Pan, J. Understanding factors influencing Geographic Variation in Healthcare expenditures: a small areas Analysis Study. *INQUIRY: J. Health Care Organ. Provis. Financing.* **61**, 00469580231224823 (2024).
21. Rajora, K. Reviews research on applying machine learning techniques to reduce false positives for network intrusion detection systems. *Babylon. J. Mach. Learn.* **2023**, 26–30 (2023).
22. Francis, M. R. et al. Vaccination coverage and factors associated with routine childhood vaccination uptake in rural Vellore, southern India, 2017. *Vaccine.* **37**, 3078–3087 (2019).
23. Corben, P. & Leask, J. To close the childhood immunization gap, we need a richer understanding of parents' decision-making. *Hum. Vaccines Immunotherapeutics.* **12**, 3168–3176 (2016).
24. Lo Vecchio, A. et al. Determinants of low measles vaccination coverage in children living in an endemic area. *Eur. J. Pediatrics.* **178**, 243–251 (2019).
25. Riaz, A. et al. Reasons for non-vaccination and incomplete vaccinations among children in Pakistan. *Vaccine.* **36**, 5288–5293 (2018).
26. Hagemann, C., Streng, A., Kraemer, A. & Liese, J. G. Heterogeneity in coverage for measles and varicella vaccination in toddlers–analysis of factors influencing parental acceptance. *BMC Public. Health.* **17**, 1–10 (2017).
27. Mvula, H. et al. Predictors of uptake and timeliness of newly introduced pneumococcal and rotavirus vaccines, and of measles vaccine in rural Malawi: a population cohort study. *Plos One.* **11**, e0154997 (2016).
28. Visser, O. et al. Assessing determinants of the intention to accept a pertussis cocooning vaccination: a survey among Dutch parents. *Vaccine.* **34**, 4744–4751 (2016).
29. Khurana, S., Sipsma, H. L. & Caskey, R. N. HPV vaccine acceptance among adolescent males and their parents in two suburban pediatric practices. *Vaccine.* **33**, 1620–1624 (2015).
30. Biswas, S. Role of ChatGPT in Computer Programming. *Mesopotamian Journal of Computer Science.* **2023**, 9–15 (2023).
31. Hlapisi, N. M. Enhancing hybrid spectrum access in cr-iot networks: reducing sensing time in low snr environments. *Mesop. J. Comput. Sci.* **2023**, 47–52 (2023).
32. Winter, T., Riordan, B., Scarf, D. & Jose, P. Conspiracy beliefs and distrust of science predicts reluctance of vaccine uptake of politically right-wing citizens. *Vaccine.* **40**, 1896–1903 (2022).
33. Stefanizzi, P. et al. Immune response to one dose of BNT162b2 mRNA covid-19 vaccine followed by SARS-CoV-2 infection: an Italian prospective observational study. *Vaccine.* **40**, 1805–1809 (2022).
34. Kreps, S. E. & Kriner, D. L. Model uncertainty, political contestation, and public trust in science: evidence from the COVID-19 pandemic. *Sci. Adv.* **6**, eabd4563 (2020).
35. Akgün, Ö. et al. Exploring the attitudes, concerns, and knowledge regarding COVID-19 vaccine by the parents of children with rheumatic disease: cross-sectional online survey. *Vaccine.* **40**, 1829–1836 (2022).
36. Lebedev, G. Artificial intelligence in healthcare: directions of standardization. *Handb. Artif. Intell. Healthc.: 2: Pract. Prospect.* 231–257 (2022).
37. MacKay, M. et al. A review and analysis of the literature on public health emergency communication practices. *J. Community Health.* **47**, 150–162 (2022).
38. Dalum Hansen, N., Lioma, C. & Mølbak, K. in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.* 1953–1956.
39. Yang, C. C. Explainable artificial intelligence for predictive modeling in healthcare. *J. Healthc. Inf. Res.* **6**, 228–239 (2022).
40. Sabri, S. Q., Arif, J. Y. & Çınar, A. A. Comparative study of Chest Radiographs and Detection of The Covid 19 Virus Using Machine Learning Algorithm. *Mesop. J. Comput. Sci.* 34–43 (2024).
41. Raza, A. et al. AIPs-SnTCN: Predicting anti-inflammatory peptides using fastText and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks. *J. Chem. Inf. Model.* **63**, 6537–6554 (2023).
42. Akbar, S. et al. Prediction of amyloid proteins using embedded evolutionary & ensemble feature selection based descriptors with eXtreme gradient boosting model. *IEEE Access.* **11**, 39024–39036 (2023).
43. Akbar, S., Zou, Q., Raza, A. & Alarfaj, F. K. iAFPs-Mv-BiTCN: Predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks. *Artif. Intell. Med.* **151**, 102860 (2024).
44. Akbar, S., Raza, A. & Zou, Q. Deepstacked-AVPs: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model. *BMC Bioinform.* **25**, 102 (2024).
45. Raza, A., Alam, W., Khan, S., Tahir, M. & Chong, K. T. iPro-TCN: prediction of DNA promoters recognition and their strength using temporal convolutional network. *IEEE Access.* **11**, 66113–66121 (2023).
46. Ullah, M., Akbar, S., Raza, A. & Zou, Q. DeepAVP-TPPred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm. *Bioinformatics.* **40**, btae305 (2024).
47. Raza, A. et al. Comprehensive analysis of computational methods for Predicting anti-inflammatory peptides. *Arch. Comput. Methods Eng.* **31**, 3211–3229 (2024).
48. Akbar, S., Hayat, M., Kabir, M. & Iqbal, M. iAFP-gap-SMOTE: an efficient feature extraction scheme gapped dipeptide composition is coupled with an oversampling technique for identification of antifreeze proteins. *Lett. Org. Chem.* **16**, 294–302 (2019).
49. Akbar, S., Rahman, A. U., Hayat, M., Sohail, M. & cACP Classifying anticancer peptides using discriminative intelligent model via Chou's 5-step rules and general pseudo components. *Chemometr. Intell. Lab. Syst.* **196**, 103912 (2020).
50. Ali, F. et al. AFP-CMBPred: computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. *Comput. Biol. Med.* **139**, 105006 (2021).
51. Ali, F., Ahmed, S., Swati, Z. N. K. & Akbar, S. DP-BINDER: machine learning model for prediction of DNA-binding proteins by fusing evolutionary and physicochemical information. *J. Comput. Aided Mol. Des.* **33**, 645–658 (2019).
52. Akbar, S. et al. Identifying neuropeptides via evolutionary and sequential based multi-perspective descriptors by incorporation with ensemble classification strategy. *IEEE Access.* **11**, 49024–49034 (2023).
53. Akbar, S. et al. iAtbP-Hyb-EnC: prediction of antitubercular peptides via heterogeneous feature representation and genetic algorithm based ensemble learning model. *Comput. Biol. Med.* **137**, 104778 (2021).
54. Wang, D., Willis, D. R. & Yih, Y. The pneumonia severity index: Assessment and comparison to popular machine learning classifiers. *Int. J. Med. Informatics.* **163**, 104778. https://doi.org/10.1016/j.ijmedinf.2022.104778 (2022).
55. Akbar, S. et al. Prediction of antiviral peptides using transform evolutionary & SHAP analysis based descriptors by incorporation with ensemble learning strategy. *Chemometr. Intell. Lab. Syst.* **230**, 104682 (2022).
56. Ahmad, A., Akbar, S., Tahir, M., Hayat, M. & Ali, F. iAFPs-EnC-GA: identifying antifungal peptides using sequential and evolutionary descriptors based multi-information fusion and ensemble learning approach. *Chemometr. Intell. Lab. Syst.* **222**, 104516 (2022).

57. Rukh, G., Akbar, S., Rehman, G., Alarfaj, F. K. & Zou, Q. StackedEnC-AOP: prediction of antioxidant proteins using transform evolutionary and sequential features based multi-scale vector with stacked ensemble learning. *BMC Bioinform.* **25**, 256. https://doi.org/10.1186/s12859-024-05884-6 (2024).
58. Lee, S., Riley-Behringer, M., Rose, J. C., Meropol, S. B. & Lazebnik, R. Parental vaccine acceptance: a logistic regression model using previsit decisions. *Clin. Pediatr.* **56**, 716–722 (2017).

Acknowledgements
The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/455/45.

Author contributions
MB.Q performed Supervision, Idea, Writing, and Proof-Reading.MS.Q performed Writing, formal Analysis, and Methodology.U. I performed Writing, methodology, implementation, and visualization. A.R performed Writing, Model Creation, methodology, and Validation. Y.G performed Writing, Formal Analysis, and Validation. N.I performed Writing, Data curation, and interpretation.M. A Performed writing, software, visualization, and Model curation. A. Q performed visualization, writing, data curation, and Funding.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to A.R. or A.Q.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.