DOI: 10.15849/IJASCA.250730.02

Int. J. Advance Soft Compu. Appl, Vol. 17, No. 2, July 2025
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Data-Driven Insights into Vaccine Hesitancy: A Machine Learning Analysis of Health, Lifestyle, and Socioeconomic Factors

Mohammad Kharabsheh[1], Ali Alsarhan[2], Nadera Aljawabrah[3], Hind Milhem[4], Hebah Alquran[5], and Shadi Banitaan[6]

[1, 3,4]Department of Computer Information Systems, Faculty of Prince Al-Hussein bin Abdullah II of Information Technology, The Hashemite University, Zarqa, Jordan.
e-mail: mohkh86@hu.edu.jo, naderam@hu.edu.jo, hindais@hu.edu.jo
[2]Department of Computer Information Systems, The University of Maryland, Baltimore County, Baltimore, USA.
e-mail: ali.alsarhan@umbc.edu
[5]Faculty of Information Technology and Computer Sciences, Yarmouk University, Irbid, Jordan.
e-mail: hebah.alquran@yu.edu.jo
[6]Department of Electrical and Computer Engineering and Computer Science, College of Engineering and Science, University of Detroit Mercy, Detroit, USA.
e-mail: banitash@udmercy.edu

Abstract

The persistent challenge of COVID-19 vaccine hesitancy hinders global efforts to achieve herd immunity. As a public health barrier, vaccine hesitancy diminishes the impact of immunization programs and prolongs population-level vulnerability. This study investigates vaccine acceptance factors by developing a predictive model based on sociodemographic, health, and lifestyle variables, including smoking status. Data were collected through a structured survey of 500 participants representing diverse demographic backgrounds. The survey included questions on demographics, smoking status, prior COVID-19 infection, health conditions, and attitudes toward vaccine safety. We employed Google Cloud's Vertex AI AutoML to train and evaluate multiple machine learning classification algorithms. Random Forest and Support Vector Machines (SVM) achieved the highest predictive performance among these. The final model demonstrated strong classification accuracy (93%) and a high AUC score (0.96), underscoring its robustness. Feature importance analysis revealed that individuals concerned about long-term vaccine safety were 2.5 times more likely to be vaccine-hesitant. The perception of low personal risk from COVID-19 was also a major contributing factor. By contrast, lifestyle variables such as smoking status had a comparatively weaker association with hesitancy.

This study contributes to the growing application of machine learning in public health by presenting a scalable, interpretable framework for identifying populations at higher risk of vaccine hesitancy. These findings provide actionable insights for health authorities, emphasizing the need for communication strategies that directly address safety concerns and risk misperception. Tailored outreach should prioritize individuals with lower educational attainment, where hesitancy was notably more prevalent. These

Received 25 March 2025; Accepted 2 May 2025

contributions offer a foundation for more effective vaccine campaigns and broader pandemic response efforts.

Keywords: *COVID-19, Vaccine Hesitancy, Machine Learning, Predictive Modeling, Feature Importance, Public Health, Survey Research.*

1 Introduction

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) which enables machines to learn from data and continuously improve their performance without the need for explicit programming for each individual scenario. ML is primarily concerned with designing systems which can learn from dataset and make decisions or predictions based on that knowledge [1]. At the core of ML is the development of algorithms capable of learning through training on diverse inputs [2, 3]. One of the most widely used approaches in ML is supervised learning, where the algorithm learns from input-output pairs and seeks to map inputs (features) to their correct outputs (labels). The aim of supervised learning is to identify patterns in the distribution of class labels based on the in- put features and apply these patterns to classify new, unseen instances [4, 5]. Given that datasets often contain many features, some of which may be less informative, Feature Selection (FS), or attribute selection, becomes a crucial step in developing a robust and effective predictive model [6, 7].

During the COVID-19 pandemic, ML models have been used to forecast patient outcomes, monitor disease spread, and support clinical decision-making. However, limited research applies ML to behavioral and social aspects of the pandemic, such as vaccine hesitancy. Vaccine hesitancy is a growing global concern that reduces the effectiveness of immunization campaigns and endangers community-level health protection. Understanding and predicting vaccine hesitancy is critical to guide public health outreach and communication strategies, particularly in communities with lower vaccine uptake. Undoubtedly, the COVID-19 pandemic has fundamentally transformed the world's perspective on life.

This global health crisis has called for unprecedented efforts and collaboration, making it essential to utilize the best available technologies to guide public health decisions and responses [8, 9]. The coronavirus, SARS-CoV-2, continues to posture an important global health challenge [10]. Although researchers have made rapid progress in developing evidence around pharmaceutical treatments, the absence of perfect established preventive measures has complicated the efficient triage the patients of COVID-19. While tools like the Modified Early Warning Score (MEWS) [11] are valuable for assessing the severity of illness in COVID-19 patients, there is limited research examining the ability of these scoring systems to predict patient outcomes, including mortality [12]. Recent studies have highlighted that the discriminatory power of such rule-based scoring systems often lacks precision and quality [10]. From an epidemiological perspective, it is anticipated that hospitals will experience a surge in COVID-19 patient admissions. Health systems are working to maintain sustainable triage processes to optimize the distribution of limited resources [13]. Early identification of patients at risk of deterioration those likely to require mechanical ventilation could allow physicians to monitor these individuals more closely, thereby creating a more manage environment for intubation. Delays in intervention and subsequent urgent intubation of serious ill patients are associated with significant risks, including peri-intubation hypoxia, hypotension, arrhythmias, and even cardiac arrest.

In contrast to these clinical applications of ML, this study focuses on understanding behavioral patterns related to COVID-19 vaccine hesitancy using supervised classification

models. Specifically, we aim to identify the most influential sociodemographic, health, and lifestyle features contributing to an individual's decision to accept or reject vaccination. The scope of this research is limited to analyzing responses collected through a structured survey administered to a general population sample. The survey included questions about vaccine attitudes, prior infection, education, health status, and lifestyle habits such as smoking. We developed a predictive model by applying Google Cloud's Vertex AI AutoML platform and extracted feature importance scores to inform future public health messaging and intervention strategies.

The COVID-19 pandemic has considered unprecedented pressure on healthcare systems globally, focusing the urgent need for advanced tools to tackle this ongoing public health and clinical crisis. Pneumonia commonly develops during the third week of symptomatic infection, with a mortality rate ranging from 3-10%, which significantly raises the risk of multi-organ failure and necessitates mechanical ventilation. Patients of- ten report a sudden onset of dyspnea during routine activities or at rest [15, 16]. Key clinical indicators include a respiratory rate exceeding 30 breaths per minute, blood oxygen saturation 93%, and a PaO2/FiO2 ratio below 300 mmHg. These signs indicate the early stages of Acute Respiratory Distress Syndrome (ARDS), which can escalate from mild to severe respiratory failure. Despite these indicators, there remains clinical uncertainty about the progression of a patient's condition and the optimal timing for initiating mechanical ventilation in cases of respiratory failure. Machine learning models, as an example the one developed in this paper, show promise in generating predictive tools that assist in clinical decision-making across various outcomes. Recently, such models have been applied during the COVID-19 crisis to support clinical assessments [17].

As governments and health organizations work to mitigate the impact of the COVID-19 pandemic, data science and machine learning technologies have played a critical role in supporting these efforts, particularly in tracking transmission trends, aiding diagnosis, optimizing disinfection protocols, and accelerating vaccine development. So far, machine learning and data science have proven to be top of most powerful tools in battling the range of the virus, playing a crucial role in helping China curb its transmission in record time [18, 19].

Pandemics and infectious diseases have been major concerns for many decades. One major solution was the development of effective vaccines that have been rolled out globally, such as those from Pfizer- BioNTech, Moderna, etc.

This research aims to contribute to this evolving area by applying automated machine-learning techniques to model vaccine hesitancy and interpret feature importance. By identifying key hesitancy predictors, this study provides findings that can help public health authorities develop focused communication strategies tailored to specific population segments.

Machine learning can be used to better understand COVID-19 vaccine hesitancy and to provide a predictive model based on health, lifestyle, and Socioeconomic Variables.

Furthermore, using visualization, we highlight the most common variables that show vaccine aversion. We show that feature importance analysis provides critical insights into the key factors influencing vaccination decisions, enabling public health authorities to design targeted strategies to address vaccine hesitancy.

In [48], The researchers evaluated the performance of various ma- chine learning and deep learning models in detecting vaccine-hesitant tweets during the COVID-19 pandemic. The results show that deep learning methods, specifically Long Short-Term Memory (LSTM)

and Recurrent Neural Networks (RNN), outperformed traditional machine learning models, achieving an accuracy of 86% compared to 83%. These findings highlight the effectiveness of deep learning approaches in analyzing social media content for public health insights.

The study employed supervised classifiers by grouping the dataset into two categories: (1) training and (2) testing. We trained the ma- chine learning models in the training group and evaluated their performance using the testing group. We applied a 10-fold cross-validation technique [2, 20, 21, 22] to generate unbiased training and testing sets. Finally, we developed decision support models using a set of classification algorithms commonly applied in healthcare industry contexts [23, 24, 25, 26].

The structure of this paper is as follows: In section 2 provides a literature review, Section 3 outlines the methodology employed in this paper, Section 4 presents the experimental results and discussion, Section 5 offers practical recommendations, Section 6 addresses the primary threats to validity, and Section 7 conclusion with suggestions for future research.

2 Related Work

Burdick et al. [14] conducted a study aimed at enhancing machine learning (ML)-based models for predicting serious illness outcomes in COVID-19 patients. The objective was to assess Machine Learning driven risk prediction models could aid in managing the patients of COVID-19 in clinical settings. The READY clinical trial (" Respiratory Decompensation and Pattern for the Triage of COVID-19 Patients: A Prospective Study") enrolled 197 patients. The findings revealed that the algorithm outperformed the traditional early warning system, the Modified Early Warning Score (MEWS) [27], with a higher diagnostic odds ratio (DOR = 12.58) for predicting ventilation needs. Further- more, the algorithm demonstrated significantly higher sensitivity (0.90) compared to MEWS (0.78), while maintaining a high specificity ($p < 0.05$).

Patanavanich and Glantz [28] conducted a meta-analysis investigating the link between smoking and the progression of COVID-19. The results indicated that smoking is a significant risk factor for disease progression, with smokers exhibiting higher odds of progression compared to non-smokers.

Ferrari et al. [17] estimated a 48-hour prognosis for patients with mild to acute respiratory failure requiring mechanical ventilation. The study, which involved 198 patients and generated 1,068 usable observations, led to the development of three predictive models based on clinical symptoms, laboratory biomarkers, and a combination of both. The final boosted mixed model, which included 20 selected variables from the combined model, achieved the highest predictive performance (AUC = 0.84). This model demonstrated 84% accuracy in prognosis and was considered valuable for supporting clinical decision-making and the development of high-readiness-level analytics tools.

Lyu et al. [29] employed both qualitative and quantitative chest CT indicators to evaluate the clinical severity of COVID-19 pneumonia and identify the topographic features of severe cases. The study included 51 patients with COVID-19 pneumonia, categorized into three groups: normal cases (Group A, n = 12), severe cases (Group B, n = 15), and critical cases (Group C, n = 24). CT findings were analyzed using various statistical tests and ROC analysis. The results showed that, as the severity of the disease increased, there were more affected lung lobes and segments, as well as higher frequencies of consolidation, ground-glass opacities, and the" crazy-paving" pattern. Qualitative indicators, such as the total lung severity score, consolidation, and crazy-paving patterns, were effective in

distinguishing severe and critical cases from normal cases, achieving sensitivity of 69%, specificity of 83%, and ac- curacy of 73%. However, these indicators showed similar results between the severe and critical groups. When combined, the indicators demonstrated high performance, with sensitivity rates of 90% and 92%, specificity rates of 100% and 87%, and accuracy rates of 92% and 90%, respectively. Critical cases had higher overall severity scores (≥10) and greater consolidation scores (≥4) compared to normal cases.

Alshirah and Al-Fawa'reh [30] focused on detecting phishing URLs using ML-based lexical feature analysis. The study extracted lexical features from URLs and used them as inputs to various machine learning classifiers, including Random Forest (RF), Decision Tree (DT), Gaussian Naive Bayes (GNB), k-Nearest Neighbors (KNN), Logistic Regression, Support Vector Classifier (SVC), Quadratic Discriminant Analysis (QDA), Perceptron, and SMOTE. The dataset included four attack types: Defacement, Spam, Phishing, and Malware, with a focus on phishing. Among all the tested models, the Random Forest classifier achieved the highest accuracy (98%) and the best precision and recall scores (both 98%).

Although the above studies demonstrate the versatility and performance of ML models in medical and non-medical domains, relatively few have applied ML to model COVID-19 vaccine hesitancy specifically. One such study by Nyawa et al. [48] evaluated the use of deep learning models (e.g., LSTM and RNN) to detect vaccine-hesitant tweets during the COVID-19 pandemic. While their work accurately identified vaccine-related attitudes from text (up to 86%), it focused on social media data and did not investigate structured, individual-level features. In contrast, the current study addresses this gap by applying machine learning to a structured dataset obtained from a general population survey. It aims to predict vaccine hesitancy based on sociodemographic, health, and lifestyle variables such as education level, smoking status, prior COVID-19 infection, and vaccine safety concerns. This study's scope centers on individual-level modeling rather than content analysis, and its contribution lies in using feature attribution to identify and rank the key predictors of hesitancy. This research offers actionable insights for public health communication strategies by clearly defining the predictors and applying interpretable models.

3 Methodology

In this study, Google Cloud's Vertex AI AutoML was employed to develop and evaluate a tabular classification model, with the primary goal of analyzing feature importance for a previously unclassified dataset.

This research specifically aims to predict COVID-19 vaccine hesitancy based on individual-level sociodemographic, health, and lifestyle variables. The scope is limited to the analysis of self-reported data collected through a structured survey conducted with 500 participants. The study does not investigate longitudinal trends, geographic clustering, or external behavioral data such as social media. Rather, it focuses on identifying statistically significant predictors of vaccine acceptance using supervised machine learning methods.

The dataset, data-final, was processed in Google Cloud to identify key features, with a class label designated as the target variable. The data was split into training, validation, and testing subsets using an

80/10/10 partition. Leveraging AutoML's capabilities, the model training process, executed in the us-central1 region, utilized Google-managed encryption and hyper

parameter tuning to optimize the model for AUC ROC, ensuring high discriminative power. The AutoML pipeline automated the end-to-end workflow using a server less architecture, integrating tasks such as model initialization, training, and finalization. The training process, which lasted approximately 2 hours and 28 minutes, employed Shapley sampling for feature attribution, providing in- sights into the contribution of individual features to the model's pre- dictions. High-performance evaluation metrics were achieved, including a PR AUC of 0.991, ROC AUC of 0.993, and an F1 score of 0.991 at a threshold of 0.5, confirming the model's robustness in precision and recall. The pipeline execution, monitored through labels and debugging features, ensured efficient orchestration and scalability. This approach highlights the effectiveness of Google Cloud's Vertex AI AutoML in automating complex workflows, enabling the extraction of meaningful insights from tabular data through advanced feature attribution and performance evaluation techniques.

This section presents the methodology used in our study, as shown in Figure 1. We begin by detailing the data that used for assist, with an emphasis on the data collection and processing procedures. Next, we introduce the factors involved in training the classifier. Finally, we describe the model that we are developed and the metrics which used in the experimental results.

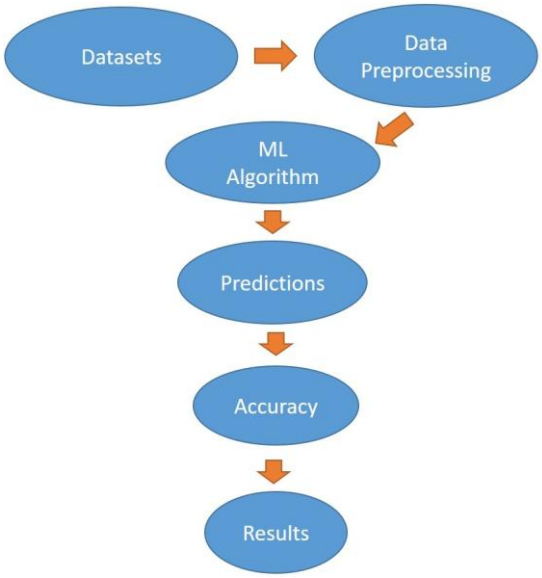

Figure 1: The Proposed Methodology.

3.1 Studied Dataset

This study was conducted through a survey. This study investigates the factors influencing vaccine acceptance by developing a predictive model based on sociodemographic, health, and lifestyle variables, including smoking status.

Table 1 outlines the classification factors utilized in the survey. These factors include personal information such as age, gender (i.e. male or female), education level (i.e. less than high school, high school, bachelor's degree, master's degree, etc.), and country, which were used to formulate the survey questions. Additional questions focused on factors like smoking status (" Are you a smoker?"), prior COVID-19 infection (" Have you had COVID-19 before?"), and health conditions (" Do you have any of the following conditions?"). Some questions required responses on a frequency scale, such as" never,"

”rarely,” ”sometimes,” or ”often,” while others were simple ”Yes” or ”No” questions. The survey also included questions about side effects experienced after receiving the COVID-19 vaccine and its safety, as well as statements where participants indicated agreement or disagreement.

Table 1: Outline of Classification Factors

No	Classification Factors
1	Personal information
2	Smoking status
3	Prior COVID-19 infection
4	Health conditions
5	Frequency scale questions
6	”Yes” or ”No” questions
7	Side effects questions experienced after receiving the COVID-19 vaccine
8	vaccine safety
9	”Agree” or ”Disagree” questions

3.2 Creating the Corpus

The crucial step in performing the classification is to generate the corpus that defines the input for the classifiers.

3.3 Classification Algorithms

In our study, we use supervised classifiers, where the dataset is split into two sets: a training set and a testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate the model’s performance. To ensure unbiased results, we applied the widely used 10-fold cross-validation technique [31]. This approach not only provides more reliable performance metrics but also yielded improved results on our dataset.

Various classification algorithms are commonly utilized in decision support systems within the healthcare studies and have been applied to develop the models that used in [32, 33], these algorithms are as follows:

- *Support Vector Machine (SVM)*: Support Vector Machines (SVM) aim to determine the optimal decision boundary between classes by maximizing the margin between them. One limitation of this approach is that it is primarily suited for binary classification tasks [34]. SVM constructs an optimal separating line (or hyperplane) that maximizes the distance between the closest data points of different classes, known as support vectors [35]. Additionally, it projects training instances into a higher-dimensional space to make the data linearly separable. This algorithm is widely used due to its effectiveness in high-dimensional spaces, often resulting in improved classification accuracy. [36].
- *Random Tree*: consider an ensemble training technique used for classification. It consists of multiple independent decision trees, each built from various samples and subsets of the training data. As a supervised learning algorithm, it extracts a random subset of data to construct each decision tree, producing a collection of individual learners. Random Tree can handle both classification and regression tasks and operates as a group of tree-based predictors, often referred to as a forest. During classification, the input feature vector is passed through each tree in the ensemble, and the final prediction is based on the combined outputs. This algorithm is particularly effective for large-scale data mining tasks, as it leverages multiple decision trees to improve accuracy and robustness [37, 38].

- *Decision Tree*: particularly, the J48 algorithm is widely used for classifying various datasets and often yields accurate classification results. It is considered one of the most effective machine learning algorithms for continuously analyzing and categorizing data. However, J48 tends to consume more memory, which can negatively impact both performance and the accuracy of the classification. The algorithm constructs a binary decision tree for classification tasks, splitting the data into ranges based on attribute values identified in the training dataset. [39].
- *Naive Bayes*: is a simple but powerful classification algorithm based on Bayes' Theorem, which provides a probabilistic frame- work for classification. It's called" naive" because it makes the assumption that all features (attributes) are independent, which often doesn't hold in practice. Despite this simplifying assumption, Naive Bayes performs well in many real-world applications, especially for text classification and spam detection [40].
- *Sequential Minimal Optimization (SMO)*: is an efficient algorithm designed to address the quadratic programming (QP) problem that arises during the training of Support Vector Machines (SVMs). Developed by John Platt in 1998, SMO decomposes the large QP problem into a series of smaller, analytically solvable sub-problems, significantly reducing computational complexity and memory requirements. SMO is widely used for training SVMs due to its high-speed training capabilities. The algorithm employs polynomial kernels to transform input features into higher-dimensional spaces, enabling the modeling of non-linear relationships. Additionally, it handles nominal attributes by mapping them to bi- nary values, facilitating their integration into the SVM framework. This approach enhances the algorithm's flexibility and applicability across various data types. [41].
- *Logistic Regression*: is a predictive analysis technique used to estimate the probability of a dependent variable based on one or more independent variables. Despite its name, it is a linear model for classification rather than regression. This approach employs a logistic function to model the posterior class probabilities for each of the required classes in the dataset. By transforming the output of a linear combination of input features into a probability value between 0 and 1, logistic regression facilitates the assignment of observations to discrete categories (e.g., 0 or 1) [42].
- *K -Star*: is a non-parametric, instance-based classification algorithm similar to the k-Nearest Neighbors (k-NN) algorithm. How- ever, it incorporates an enhanced method of calculating distances between instances using entropy-based measures rather than just Euclidean distance. K-Star is particularly useful for handling categorical data. [43].
- *Decision Table*: is a predictive method derived from decision trees, consisting of an ordered set of If-Then rules. It is often considered more efficient and simpler than traditional decision trees. One of the key advantages of decision tables is that they provide an easier, less computationally intensive alternative to decision tree-based algorithms. A Decision Table Classifier is created by using best-first search and can incorporate cross-validation for model evaluation through estimating feature subsets. To generate the decision table for a given dataset, a grouping-and-counting technique is applied, which helps classify an unknown sample based on the established rules from the table. [44].
- *K -Nearest Neighbor (K-NN)*: is a non-parametric algorithm used for both classification and regression tasks. It is particularly popular as a text classification method due to its simplicity and effectiveness. K-NN's learning phase involves storing all training

instances and deferring the decision on how to generalize the data until a new instance is encountered. This characteristic earns it the label of a lazy learner. The algorithm classifies new instances by identifying the closest known instances (i.e., nearest neighbors) and then applying a majority voting approach to determine the class of the unknown instance. [45].

- *IBk*: is a nearest-neighbor algorithm that utilizes distance metrics derived from the training set to identify the closest matching vectors, to classify data instances in the testing set [46].

4 Results and Discussions

To assess the effectiveness of our classification model, we evaluated it using the following metrics:

- *Precision*: This measures the proportion of retrieved instances that are truly relevant. It is calculated as (P = True Positives / (True Positives + False Positives)) [47].
- *Recall*: This metric evaluates the proportion of relevant instances that were retrieved by the classifier. It is computed as (R= True Positives / (True Positives + False Negatives)) [47].
- *F -Measure*: This metric combines both precision and recall to provide a single score that balances the two. It is calculated as ((2 * Recall * Precision) / (Recall + Precision)). The value of this metric is between 0 and 1 [47].
- *Accuracy*: This represents the proportion of correctly classified in- stances (both positive and negative) and is calculated as (R= (True Positives + True Negative) / (True Positives + True Negative + False Negatives + False Positive)) [47].

Figure 2 shows the relationship between education level and vaccine hesitancy, focusing on the average hesitancy rate across different education groups. The x-axis lists education levels, from" Less than High School" to" Graduate Degree," while the y-axis represents the hesitancy rate as a percentage. From the chart:

- People with less than a high school education and high school education report the highest vaccine hesitancy, both exceeding 25
- Hesitancy rates drop significantly for those with higher education levels. For instance:

– Degrees show around 10
– Bachelor's degrees drop further to below 10
– Graduate degrees report the lowest hesitancy rates, nearly 0

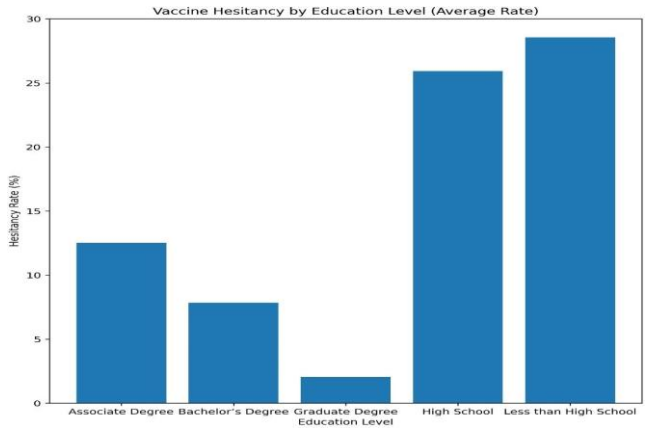

Figure 2: Vaccine Hesitancy by Education Level (Average Rate)

This visualization highlights how education level influences vaccine hesitancy, with higher education levels correlating to lower hesitancy. It provides evidence for targeted interventions, suggesting that education- focused strategies could help address vaccine hesitancy in populations with lower education levels.
Figure 3 reveals whether males or females have higher vaccine hesitancy rates, providing insights into gender-based differences.
Figure 4 compares vaccine hesitancy rates between genders, pro- viding insights into gender-related differences in hesitancy. The x-axis shows the two gender groups, Female and Male, while the y-axis represents the hesitancy rate as a percentage. From the chart:

- Females have a noticeably higher hesitancy rate compared to males, reaching around 20.
- Males show a significantly lower hesitancy rate, barely exceeding 5.

This visualization highlights a clear gender gap in vaccine hesitancy, with females reporting hesitancy at a much higher rate than males. This difference suggests that addressing vaccine concerns among females could be a critical focus for improving overall vaccination rates. Tailored communication and outreach strategies might help reduce hesitancy in this group.

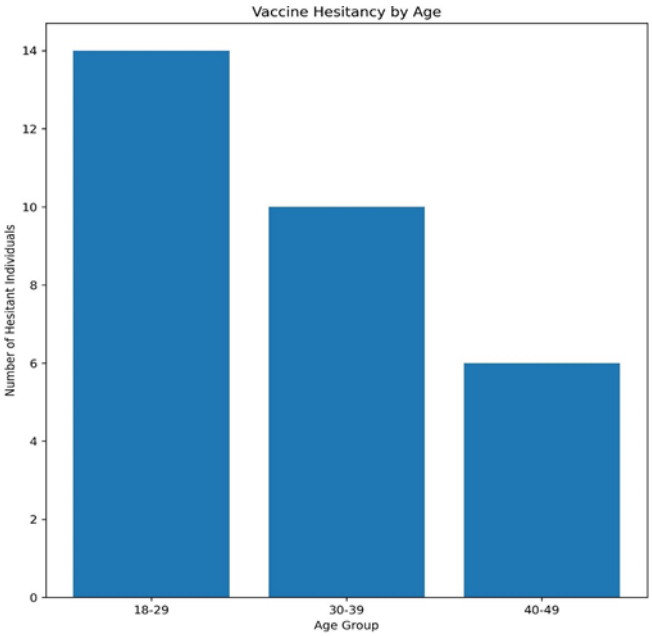

Figure 3: Vaccine Hesitancy by Age

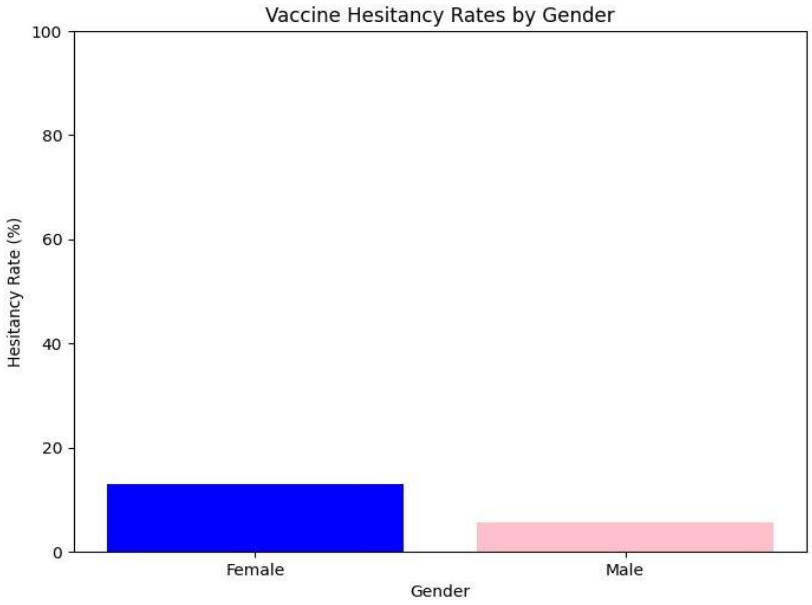

Figure 4: Vaccine Hesitancy by Gender

Figure 5 shows the feature importance analysis provides critical in- sights into the key factors influencing vaccination decisions, enabling public health authorities to design targeted strategies to address vaccine hesitancy. By identifying which variables such as age, education level, or beliefs about vaccine safety and effectiveness are most strongly associated with vaccination status, policymakers can focus efforts on specific demographics or beliefs that contribute to hesitancy. For instance, if the model highlights vaccine safety concerns as a major factor, educational campaigns can be tailored to address these fears with data-driven reassurance. Similarly, if certain age groups or educational levels show higher hesitancy, resources can be allocated effectively to engage those populations. This approach ensures that interventions are both evidence-based and strategically focused, maximizing their impact in increasing vaccination rates.

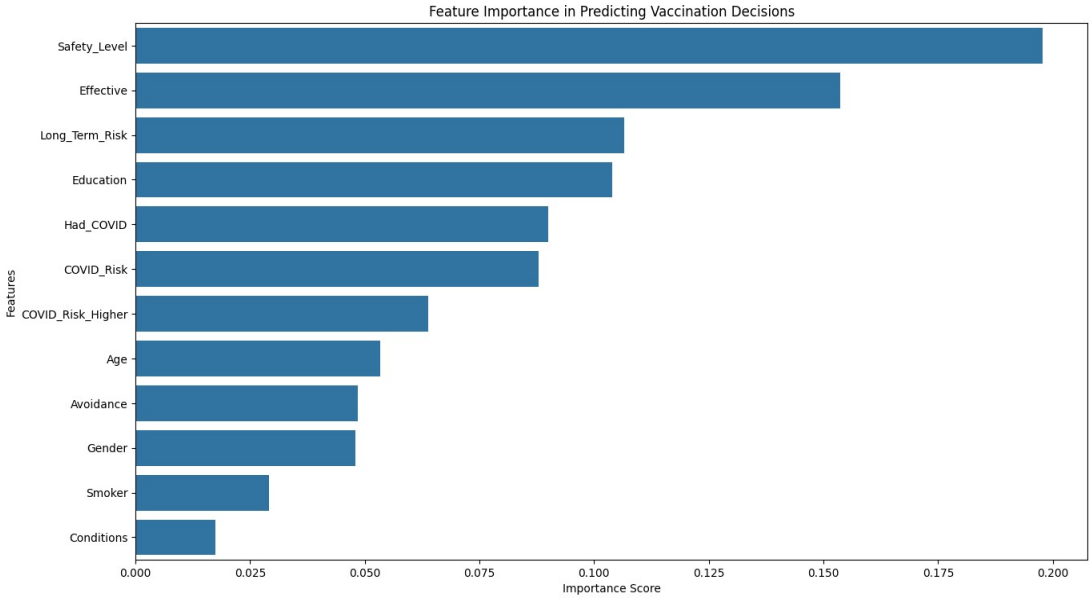

Figure 5: Feature Importance in Predicting Vaccination Decisions

In Figures 6 and 7 we give a clear picture of how well the model predicts vaccination decisions. The accuracy bar chart shows how accurate the model is overall, with the score ranging from 0 to 1, making it easy to see how well the model performs. The classification report heatmap breaks things down further, showing precision, recall, and F1 scores for each class. Darker colors mean better performance, so you can quickly spot where the model is doing well or struggling. Finally, the confusion matrix gives a detailed look at where the predictions went right or wrong, helping to understand specific patterns or mistakes. Together, these visuals make it simple to evaluate the model's performance and identify areas for improvement.

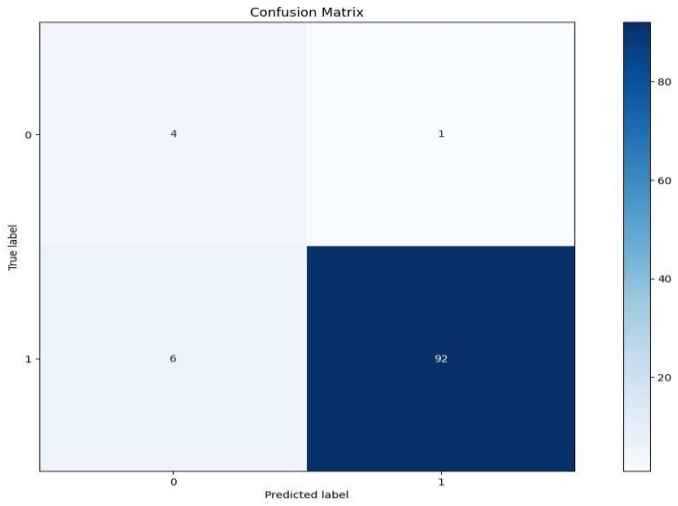

Figure 6: Confusion Matrix

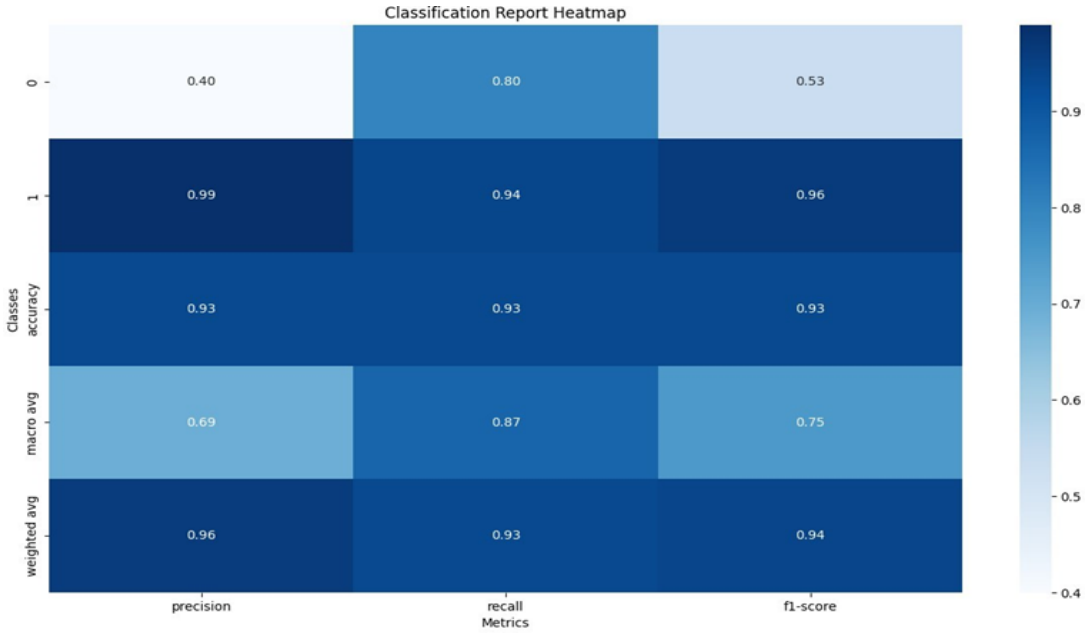

Figure 7: Classification Report Heatmap

Figures 8, 9 and 10, these charts are all about understanding how different people experience vaccine side effects. The bar chart gives a quick look at how common each severity level is, from" None" to" Severe" and beyond. The stacked bar charts dive a bit deeper, showing how side effects vary across age groups and between genders. The main goal here is to spot trends and patterns. For example, are younger people more likely to report mild side effects? Are severe reactions more common in a specific group? These findings can inform targeted strategies for vaccine rollouts, particularly by identifying specific demographic groups that may benefit from tailored communication or support. It's all about turning this data into insights that make the whole process safer and smoother for everyone.

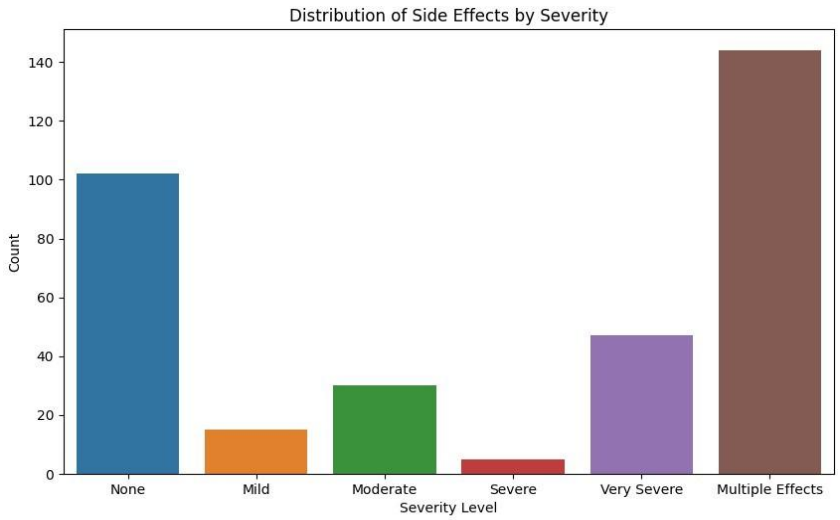

Figure 8: Distribution of Side Effect by Severity

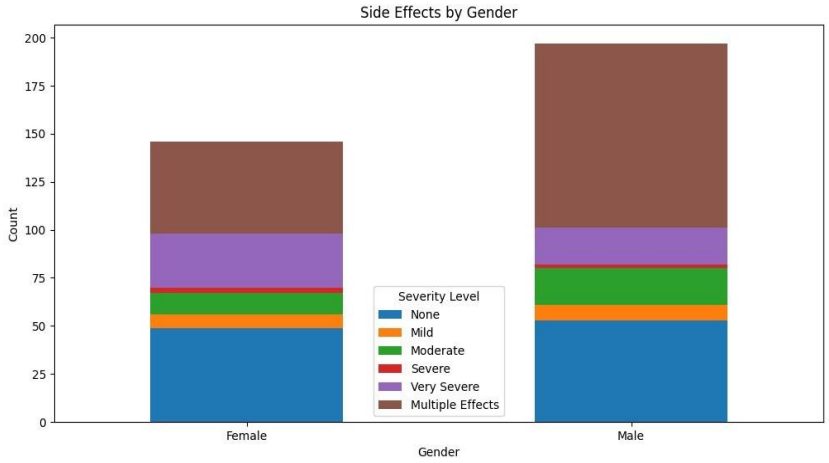

Figure 9: Side Effects by Gender

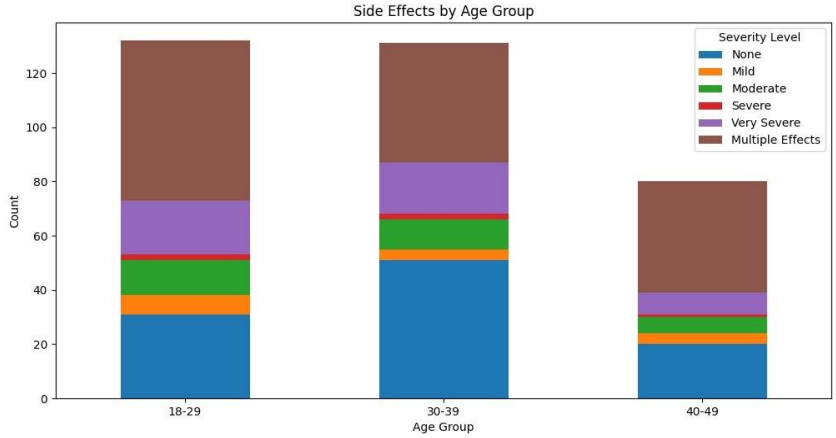

Figure 10: Side Effects by Age Group

The heatmap in Figure 11 shows how males and females experience vaccine side effects differently. The x-axis represents the side effect levels (e.g., None, Mild, Moderate, Severe), and the y-axis separates males and females. The darker the color, the more people reported that side effect level. Observations:

- None (Level 1) is the most common side effect, with slightly more males (53) than females (49).
- Multiple Effects (Level 6) is a standout, with males reporting it nearly twice as much as females (96 vs. 48).
- Moderate and severe side effects (Levels 3, 4, and 5) are less com- mon, and severe reactions are rare overall.
- Males report higher counts across most levels compared to females. Insights:
- Gender Differences: Males seem to report either no side effects or multiple side effects more often than females. This could be due to biological reasons or reporting habits.
- Severe Side Effects Are Rare: Both genders report very few severe reactions, which suggests the vaccine is well-tolerated.
- Public Health Focus: The higher reports of" Multiple Effects" among males could mean this group needs more follow-ups. Encouraging females to report symptoms more actively might also help balance the data.

- Research Hypotheses: Why do males report more extreme out- comes (None or Multiple)? Are females underreporting? This heatmap gives a starting point to explore these trends, providing insights for better vaccine rollout strategies and post-vaccination care.

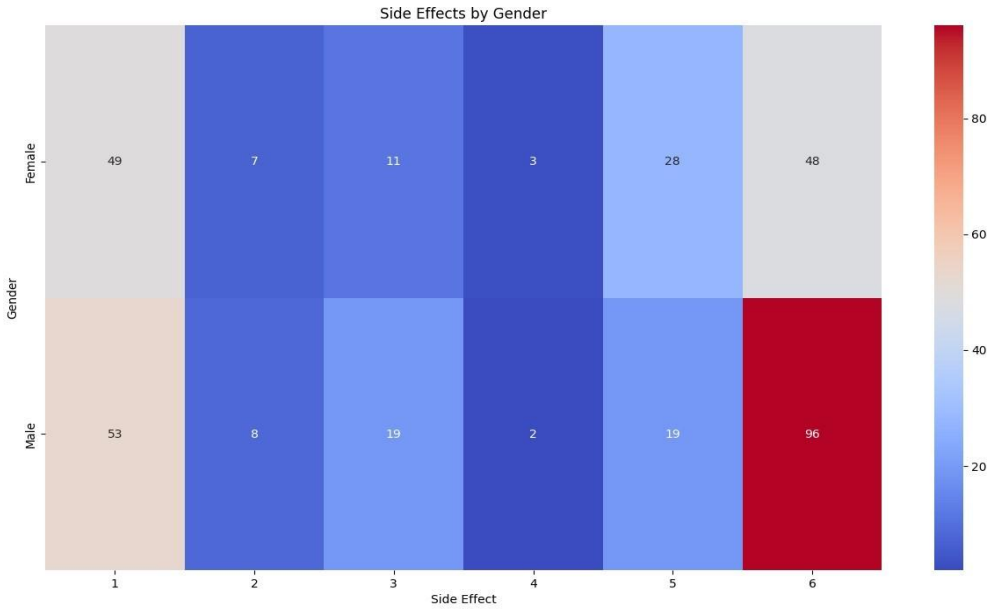

Figure 11: Side Effects by Gender

5 Recommendations

The findings of this study provide actionable insights for public health authorities seeking to address COVID-19 vaccine hesitancy. Specifically, targeted communication strategies are warranted to mitigate concerns regarding long-term vaccine safety and to rectify misperceptions of individual risk. These strategies should be evidence-based and tailored to address the primary drivers of hesitancy identified in this research.

A paramount recommendation is the development and dissemination of clear, accessible, and scientifically sound information addressing concerns about potential long-term adverse effects of COVID-19 vaccines. These communication efforts should emphasize the rigorous safety monitoring processes in place and present evidence demonstrating the favor- able risk-benefit profile of vaccination. Furthermore, messaging should be adjusted to accurately convey the risks associated with COVID-19 infection, emphasizing the potential for severe illness, hospitalization, and long-term health consequences. This is particularly crucial given the observed influence of perceived low personal risk on vaccine hesitancy.

Given the observed prevalence of vaccine hesitancy among demo- graphic groups with lower educational attainment, targeted messaging and outreach efforts are essential. Communication strategies for these groups should prioritize the use of simplified language, visual aids, and community-based channels to enhance comprehension and build trust. Active engagement with community leaders and healthcare providers may further facilitate

the dissemination of accurate information and address specific concerns within these populations. Further qualitative research is recommended to explore the specific barriers to vaccine acceptance within these communities, enabling the development of more nuanced and effective interventions.

6 Threat to Validity

As with any case study based on a sample collected through a survey, there are potential limitations that hinder the generalizability of our findings to different datasets or settings. The dataset used in this study may not be fully representative of all possible samples, which limits our ability to extend the results to other datasets. Additionally, there may be other relevant features not included in this study that could have influenced the outcomes. While the classifiers we developed are based on well-established machine learning techniques commonly used in the literature, each method has its own limitations that could potentially impact the validity of our findings.

Furthermore, while AutoML streamlines the model development process, it also introduces certain limitations. The automated hyper parameter tuning, while optimizing for performance, can obscure the specific configuration that yields the best results, making it difficult to fully understand the model's behavior. Although AutoML provides feature importance scores, the" black box" nature of some underlying algorithms can limit the depth of transparency compared to manually con- figured models, where each parameter choice and feature selection can be meticulously analyzed. To mitigate this, we focused on using well- established and interpretable machine learning algorithms within the AutoML framework and carefully analyzed the feature importance results provided by the platform.

Moving forward, we plan to explore the development of classifiers using alternative machine learning approaches to address these issues.

7 Conclusion and Future Works

This study addressed the persistent challenge of COVID-19 vaccine hesitancy by developing a predictive model to identify factors influencing vaccine acceptance. Leveraging survey data from 500 participants and machine learning techniques within Google Cloud's Vertex AI AutoML, our analysis revealed that concerns about long-term vaccine safety and perceived low personal risk of contracting COVID-19 were the strongest predictors of hesitancy, particularly among individuals with lower educational attainment.

This research was specifically designed to model individual-level vaccine hesitancy within a general population, using structured health, demographic, and risk perception variables. As such, the scope is limited to individual-level predictors and does not extend to analyzing temporal trends or cross-national comparisons.

By providing an interpretable, high-performing model for risk pro- filing and tailored outreach planning, this study contributes a practical tool for public health agencies to more effectively target and address vaccine hesitancy within specific communities. Furthermore, our findings establish a baseline for future research to incorporate additional behavioral and regional factors. To build upon these findings, future research should prioritize qualitative investigations to explore the specific barriers to vaccine acceptance within

target communities. Understanding the nuanced, community- specific reasons for hesitancy will be crucial for developing truly effective and tailored interventions, ultimately contributing to improved vaccine uptake and global herd immunity.

References

[1] F. Thung, S. Wang, D. Lo, and L. Jiang, "An empirical study of bugs in machine learning systems," in 2012 IEEE 23rd Inter- national Symposium on Software Reliability Engineering. IEEE, 2012, pp. 271– 280.

[2] F. Eibe, M. A. Hall, and I. H. Witten, "The weka workbench. online appendix for data mining: practical machine learning tools and techniques," in Morgan Kaufmann, 2016.

[3] M. Bkassiny, Y. Li, and S. K. Jayaweera, "A survey on machine- learning techniques in cognitive radios," IEEE Communications Surveys Tutorials, vol. 15, no. 3, pp. 1136–1159, 2012.

[4] A. Al-Nusirat, F. Hanandeh, M. Kharabsheh, M. Al-Ayyoub, and N. Al-dhufairi, "Dynamic detection of software defects using supervised learning techniques," International Journal of Communication Networks and Information Security, vol. 11, no. 1, pp. 185–191, 2019.

[5] J. Alzyoud, M. Kharabsheh, S. Alzyoud, and E. Alzbon, "Use of healthcare informatics applications and data for research purposes by students: Opportunities and challenges in Jordan," in 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring). IEEE, 2019, pp. 1–6.

[6] M. Kharabsheh, O. Meqdadi, M. Alabed, S. Veeranki, A. Abbadi, and S. Alzyoud, "A machine learning approach for predicting nico- tine dependence," International Journal of Advanced Computer Science and Applications, vol. 10, no. 3, pp. 179–184, 2019.

[7] M. Kharabsheh, A. Qawasmeh, O. D. Megdadi, N. Aljawabrah, R. H. Mudallal, and S. A. Alzyoud, "A critical analysis of the relationship between depression and smoking using machine learning," International Journal of Scientific Technology Research, vol. 8, pp. 22–26, 2019.

[8] A. Kumar, P. K. Gupta, and A. Srivastava, "A review of modern technologies for tackling covid-19 pandemic," Diabetes Metabolic Syndrome: Clinical Research Reviews, vol. 14, no. 4, pp. 569–573, 2020.

[9] R. Vaishya, M. Javaid, I. H. Khan, and A. Haleem, "Artificial intelligence (ai) applications for covid-19 pandemic," Diabetes Metabolic Syndrome: Clinical Research Reviews, vol. 14, no. 4, pp. 337–339, 2020.

[10] A. Lorusso, P. Calistri, A. Petrini, G. Savini, and N. Decaro, "Novel coronavirus (sars-cov-2) epidemic: a veterinary perspective," Veterinaria italiana, vol. 56, no. 1, pp. 5–10, 2020.

[11] C. P. Subbe, M. Kruger, P. Rutherford, and L. Gemmel, "Vali- dation of a modified early warning score in medical admissions," Qjm, vol. 94, no. 10, pp. 521–526, 2001.

[12] L. Y. Hsu, P. Y. Chia, and J. Lim, "The novel coronavirus (sars- cov-2) pandemic," Ann Acad Med Singap, vol. 49, no. 3, pp. 105–7, 2020.

[13] M. Kharabsheh, S. Banitaan, H. Alomari, M. Alshirah, and S. Alzyoud." Respiratory failure in covid-19 patients a comparative study of smokers to non-smokers." Indonesian Journal of Electrical Engineering and Computer Science 27, no. 2 (2022): 1127-1137.

[14] H. Burdick, C. Lam, S. Mataraso, A. Siefkas, G. Braden, R. P. Dellinger, A. McCoy, J.-L. Vincent, A. Green-Saxena, G. Barnes et al., "Prediction of respiratory

decompensation in covid-19 patients using machine learning: The ready trial," Computers in biology and medicine, vol. 124, p. 103949, 2020.
[15] C. Solinas, L. Perra, M. Aiello, E. Migliori, and N. Petrosillo, "A critical evaluation of glucocorticoids in the management of severe covid-19," Cytokine growth factor reviews, vol. 54, pp. 8–23, 2020.
[16] S. Al-Zu'bi, B. Hawashin, A. Mughaid, and T. Baker." Efficient 3D medical image segmentation algorithm over a secured multi- media network." Multimedia Tools and Applications 80 (2021): 16887-16905.
[17] D. Ferrari, J. Milic, R. Tonelli, F. Ghinelli, M. Meschiari, S. Volpi, M. Faltoni, G. Franceschi, V. Iadisernia, D. Yaacoub et al., "Machine learning in predicting respiratory failure in patients with covid-19 pneumonia challenges, strengths, and opportunities in a global health emergency," PloS one, vol. 15, no. 11, p. e0239172, 2020.
[18] S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller et al., "Leveraging data science to combat covid-19: A comprehensive review," IEEE Transactions on Artificial Intelligence, vol. 1, no. 1, pp. 85–103, 2020.
[19] L. Surya, "How government can use ai and ml to identify spreading infectious diseases," International Journal of Creative Research Thoughts (IJCRT), ISSN, pp. 2320–2882, 2018.
[20] A. Mughaid, S AlZu'bi, A. Hnaif, S. Taamneh, A. Alnajjar, and E. Abu Elsoud." An intelligent cyber security phishing detection system using deep learning techniques." Cluster Computing 25, no. 6 (2022): 3819-3828.

[21] W.-L. Zuo, Z.-Y. Wang, T. Liu, and H.-L. Chen, "Effective detection of parkinson's disease using an adaptive fuzzy k-nearest neighbour approach," Biomedical Signal Processing and Control, vol. 8, no. 4, pp. 364–373, 2013.
[22] C.-H. Jen, C.-C. Wang, B. C. Jiang, Y.-H. Chu, and M.-S. Chen, "Application of classification techniques on development an early- warning system for chronic illnesses," Expert Systems with Ap- plications, vol. 39, no. 10, pp. 8852–8858, 2012.
[23] C. Vaghela, N. Bhatt, and D. Mistry, "A survey on various classification techniques for clinical decision support system," International Journal of Computer Applications, vol. 116, no. 23, 2015.
[24] B. A. Thakkar, M. I. Hasan, and M. A. Desai, "Health care decision support system for swine flu prediction using naive bayes classifier," in 2010 International Conference on Advances in Re- cent Technologies in Communication and Computing. IEEE, 2010, pp. 101–105.
[25] M. W. Moreira, J. J. Rodrigues, V. Korotaev, J. Al-Muhtadi, and N. Kumar, "A comprehensive review on smart decision support systems for health care," IEEE Systems Journal, vol. 13, no. 3, pp. 3536–3545, 2019.
[26] B. Ozaydin, J. M. Hardin, and D. C. Chhieng, "Data mining and clinical decision support systems," in Clinical Decision Support Systems. Springer, 2016, pp. 45–68.
[27] J. Gardner-Thorpe, N. Love, J. Wrightson, S. Walsh, and N. Keeling, "The value of modified early warning score (mews) in surgical in-patients: a prospective observational study," The Annals of The Royal College of Surgeons of England, vol. 88, no. 6, pp. 571–575, 2006.
[28] R. Patanavanich and S. A. Glantz, "Smoking is associated with covid-19 progression: a meta-analysis," Nicotine and Tobacco Re- search, vol. 22, no. 9, pp. 1653–1656, 2020.

[29] P. Lyu, X. Liu, R. Zhang, L. Shi, and J. Gao, "The performance of chest ct in evaluating the clinical severity of covid-19 pneumonia: identifying critical cases based on ct characteristics," Investigative radiology, vol. 55, no. 7, pp. 412–421, 2020.
[30] QY. Quba, H. Al Qaisi, A. Althunibat, and S. AlZu'bi." Software requirements classification using machine learning algorithm's." In 2021 international conference on information technology (ICIT), pp. 685-690. IEEE, 2021.
[31] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," Journal of the American statistical association, vol. 78, no. 382, pp. 316–331, 1983.
[32] M. Shouman, T. Turner, and R. Stocker, "Applying k-nearest neighbour in diagnosing heart disease patients," International Journal of Information and Education Technology, vol. 2, no. 3, pp. 220–223, 2012.
[33] I. H. Witten, E. Frank, M. A. Hall, C. Pal, and M. DATA, "Practical machine learning tools and techniques," in DATA MINING, vol. 2, 2005, p. 4.
[34] W. Noble, "What is a support vector machine? nature biotechnology," 2006.
[35] K. P. Bennett and C. Campbell, "Support vector machines: hype or hallelujah?" ACM SIGKDD explorations newsletter, vol. 2, no. 2, pp. 1–13, 2000.
[36] P. Ahmad, S. Qamar, and S. Q. A. Rizvi, "Techniques of data mining in healthcare: a review," International Journal of Computer Applications, vol. 120, no. 15, 2015.
[37] B. Pfahringer, "Random model trees: an effective and scalable regression method," 2010.

[38] K.Wisaeng, "A comparison of different classification techniques for bank direct marketing," International Journal of Soft Computing and Engineering (IJSCE), vol. 3, no. 4, pp. 116–119, 2013.

[39] N. Saravana and D. V. Gayathri, "Performance and classification evaluation of j48 algorithm and kendall's based j48 algorithm (knj48)," Int. J. Comput. Trends Technol.(IJCTT)–Volume, vol. 59, 2018.

[40] I. Rish et al., "An empirical study of the naive bayes classifier," in IJCAI 2001 workshop on empirical methods in artificial intelligence, vol. 3, no. 22, 2001, pp. 41–46.

[41] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," 1998.
[42] E. W. Steyerberg, F. E. Harrell Jr, and P. H. Goodman, "Neural networks, logistic regression, and calibration," Medical Decision Making, vol. 18, no. 3, pp. 349–350, 1998.
[43] D. Y. Mahmood and M. A. Hussein, "Intrusion detection system based on k-star classifier and feature set reduction," International Organization of Scientific Research Journal of Computer Engineering (IOSRJCE) Vol, vol. 15, no. 5, pp. 107–112, 2013.
[44] G. Banerji and K. Saxena, "An efficient classification algorithm for real estate domain," India: International Journal of Modern Engineering Research (IJMER) www.ijmer. com, vol. 2, no. 4, 2012.
[45] F. Sebastiani, "Machine learning in automated text categorization," ACM computing surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.
[46] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," Machine learning, vol. 6, no. 1, pp. 37–66, 1991.

[47] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” arXiv preprint arXiv:2010.16061, 2020.
[48] S.Nyawa, D. Tchuente, and S. Fosso-Wamba, 2024. COVID-19 vaccine hesitancy: a social media analysis using deep learning. Annals of Operations Research, 339(1), pp.477-515.