



OPEN

Comparison of machine learning models for mucopolysaccharidosis early diagnosis using UAE medical records

Aamna AlShehhi^{1,2✉}, Hiba Alblooshi^{2,3}, Ruba Fadul¹, Natnael Tumzghi¹, Amal Al Tenaiji⁴, Mariam Al Harbi⁵ & Fatma Al-Jasmi^{2,3,6}

Rare diseases, such as Mucopolysaccharidosis (MPS), present significant challenges to the healthcare system. Some of the most critical challenges are the delay and the lack of accurate disease diagnosis. Early diagnosis of MPS is crucial, as it has the potential to significantly improve patients' response to treatment, thereby reducing the risk of complications or death. This study evaluates the performance of different machine learning (ML) models for MPS diagnosis using electronic health records (EHR) from the Abu Dhabi Health Services Company (SEHA). The retrospective cohort comprises 115 registered patients aged ≤ 19 Years old from 2004 to 2022. Using nested cross-validation, we trained different feature selection algorithms in combination with various ML algorithms and evaluated their performance with multiple evaluation metrics. Finally, the best-performing model was further interpreted using feature contributions analysis methods such as Shapley additive explanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME). We found that Naive Bayes trained on the domain expert selected features reported a superior performance with an accuracy of 0.93 (0.08), AUC of 0.96 (0.04), F1-score of 0.91 (0.1), and MCC of 0.86 (0.16). SHAP and LIME analysis that were conducted on the best-performing model highlighted key features related to dental manifestations and respiratory infections which are commonly presented in MPS patients, such as acute gingivitis, accretions on teeth, dental caries, acute pharyngitis, acute tonsillitis, and acute bronchitis. This study introduces a cost-effective screening approach for MPS disease using non-invasive EHR, which contributes to the advances in digital screening tools for the early diagnosis of rare diseases.

Keywords Rare diseases, Mucopolysaccharidosis, Diagnosis, Machine learning, Feature selection

Rare diseases are complex, chronic, life-threatening (majority), and disabling conditions^{1–3}. 50% to 75% of rare diseases target children, of whom 30% don't live to celebrate their fifth birthday^{1,4}. While 80% of rare diseases are genetic diseases, the rest are caused by external factors such as environment, infectious, or unknown^{1,5–7}. Globally, the disease is categorized as a rare disease when it affects fewer than 1 in 2,000 in the European Union or less than 200,000 in the USA^{1,3–5}. Recently, more than 7,000 rare diseases have been identified, affecting more than 350 to 400 million people worldwide^{1,4,5,8}. Out of the 7,000 cases, only 5% have an approved treatment⁹. Rare diseases impose several unique challenges on the economy and healthcare system. In 2016, the USA Healthcare Cost and Utilization Project (HCUP) reported that the overall estimated cost related to rare diseases is \$768 billion, comparable with \$880 billion for common diseases^{3,6}.

Patients living with RDs faced multiple challenges due to the lack of accurate and timely diagnosis; on average, physicians took approximately 7.6 years in the USA and 5.6 years in the U.K. to accurately diagnose the diseases, with multiple misdiagnoses throughout the way^{4,7,10}. This delay in rare disease diagnosis can be explained by several aspects, such as limitations in disease perception, limited knowledge of the disease by the primary physician due to the small size of the affected population, and unavailable diagnostic tests and

¹Department of Biomedical Engineering and Biotechnology, Khalifa University, Abu Dhabi, United Arab Emirates.

²ASPIRE Precision Medicine Research Institute, United Arab Emirates University, Abu Dhabi, United Arab Emirates. ³Department of Genetics and Genomics, United Arab Emirates University, Al Ain, United Arab Emirates.

⁴Department of Pediatrics, Sheikh Khalifa Medical City, Abu Dhabi, United Arab Emirates. ⁵Research Department, SEHA-Corporate Medical and Clinical Affairs, Abu Dhabi, United Arab Emirates. ⁶Department of Pediatrics, Tawam Hospital Division of Genetic Metabolic, Al Ain, United Arab Emirates. ✉email: aamna.alshehhi@ku.ac.ae

facilities^{6,9,10}. Unfortunately, the delay in finding the proper disease diagnosis can lead to the death of the patients after undergoing a painful journey of mistreatment and misdiagnosis¹⁰. Moreover, it is important to note that a lot of patients still live with the rare condition without being properly diagnosed¹¹.

Those challenges motivated the launch of several initiatives dedicated to early rare disease diagnosis and screening, such as the Korean Genetic Diagnosis Program for Rare Diseases (KGDP) Phase I and II⁶, International Rare Diseases Research Consortium (IRDiRC)⁷, and Alabama Genomic Health Initiative (AGHI)¹². Most of those initiatives focused on screening patients by targeting highly suspicious individuals, as recommended by their healthcare providers¹². This is followed by confirming the case using conventional methods such as genome-based or exome-based gene panel/genomic sequencing^{6,11–13}. In a similar vein and with the availability of electronic records, several works focused on utilizing the disease phenotype, image, and fluids to build disease diagnosis support systems¹⁴. Notably, one study proposed a phenotypic similarity algorithm based on calculating the one-to-all rank. The study focused on finding the patients' disease phenotype and comparing it with existing and annotated diseases⁷. Medical records or electronic health records (EHRs) are equally important in disease diagnosis as the conventional methods. Those records provide an efficient and cost-effective way for early disease diagnosis to improve patient outcomes and disease management¹⁵. Unfortunately, to date, only a limited number of studies and efforts have been allocated to utilizing EHRs in rare disease landscapes.

Given the scarcity of rare diseases and the challenges associated with their diagnoses, recent research has turned towards leveraging artificial intelligence (AI) technologies to predict RD conditions. Brasil et al.⁴ explored the potential and challenges of using AI in various aspects of rare diseases. It looked into how AI could be used for diagnosing and understanding rare diseases, developing treatments, maintaining patient registries, and managing health records. The study highlighted the significant impact AI could have on people affected by rare diseases. Although various medical decision support systems exist to guide diagnosis and medical treatment, the majority utilized statistical approaches, not AI techniques^{16,17}. AI technologies are not widely used in rare disease diagnosis due to the complexity of rare disease characteristics and the limited number of patients with similar phenotypes, disease severity, presentation, and progression¹⁸. Therefore, our study will be the first to fill the gap by training different AI models and comparing their performance using EHR data to diagnose patients with mucopolysaccharidosis (MPS), a rare metabolic disease.

Mucopolysaccharidoses (MPS) are a group of inherited inborn errors of metabolism disorders caused by different enzyme deficiencies involved in the breakdown of glycosaminoglycans. MPS is a progressive multisystem disorder with a heterogeneous spectrum of symptoms that varies based on the severity and the subcategory of MPS. The patients with MPS present with recurrent respiratory tract infections (upper respiratory tract infection, acute tonsillitis, pharyngitis, bronchiolitis, bronchitis, pneumonia) as well as recurrent otitis media. The other clinical presentations of these disorders are coarse facial features, macrocephaly, corneal clouding, inguinal or umbilical hernia, hepatosplenomegaly, valvular heart disease, dysostosis multiplex, limitation of joint movement, and gibbus deformity. In the severe form of the disease, they might present with hydrocephalus and developmental delay^{19–21}.

In recent years, research efforts have been directed toward employing AI techniques in various healthcare applications such as: disease diagnosis, drug discovery and development, precision medicine, and clinical trials. Garavand et al.²² used machine learning to build a diagnostic model for coronary artery disease (CAD) based on clinical examination features. They compared different ML models for their effectiveness in diagnosing CAD cases, highlighting the potentials of SVM and RF ML models in detecting CAD patients from clinical examination data. Moreover, Ghaderzadeh et al.²³ reviewed various studies that focus on the use of AI to address the antimicrobial resistance in the discovery and development of drugs, which showed the capabilities of AI models in recognizing new antimicrobial compounds, enhance existing drugs to tackle the antimicrobial resistance, and forecasting drug resistance.

Concerning the utilization of AI techniques in precision medicine, Pudjihartono et al.²⁴ discussed the strengths and weaknesses of different feature selection methods, namely filter, wrapper, and embedded methods, that could be used to overcome the curse of dimensionality challenge presented in the genotype data to build more accurate disease risk prediction models based on patients' genetic data, contributing to the advances of AI tools in precision medicine. Additionally, a study by Carlier et al.²⁵ applied ML approaches in designing an *in silico* clinical trial for a pediatric rare disease. They examined bone morphogenetic protein (BMP) treatment for congenital pseudarthrosis of the tibia (CPT) disease; by proposing an unsupervised ML model, namely ward hierarchical clustering, to cluster the virtual subject population into different groups based on their response to the BMP treatment along with the supervised Random Forest ML algorithm to identify the potential biomarkers in predicting the effectiveness of the therapy. Overall, AI-based methods have resolved many challenges in the field of healthcare.

Although AI has had noticeable and successful applications in healthcare, there is still a lack of research directed toward applying machine learning models—subfields of AI—for MPS early diagnosis in specific and rare diseases in general. In this study, we aim to fill the gap by implementing and comparing the performance of different machine learning models trained on de-identified and unstructured patients' diagnosis data extracted from the Abu Dhabi Health Services Company (SEHA) healthcare system for early diagnosis of MPS diseases. Furthermore, we will explain and interpret the best model internal behavior and decision using SHapley Additive exPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) methods to understand and report the features that directed the model decision and its clinical validity. The key contributions of this work can be summarized as follows.

1. This is the first study to utilize AI and EHR data for MPS early diagnosis.

2. We trained and compared the performance of nine machine learning models across five feature selection methods: four automated feature selection methods and one feature selection based on domain experts' knowledge.
3. We interpreted the best-performing model using SHAP and LIME, and clinically validated its outcomes using domain expert feedback.
4. Our work validates the applicability of machine learning for MPS diagnosis using only disease symptoms. This offers a non-invasive and cost-effective screening for MPS patients using EHR.

Results

Our cohort includes registered patients aged 2 to 19 years old at SEHA from 2004 to 2022 (Fig. 1). A total of 106 patients (37 MPS patients and 69 control) were eligible to be included in the study. For those patients, we extracted 1186 historical medical diagnoses such as dental caries on smooth surface penetrating into dentin, acute pharyngitis due to other specified organisms, epistaxis, diarrheal, obesity, etc., to train the various machine learning models. Using nested cross-validation, we trained different combinations of the ML models and feature selection algorithms. Across the five cross-validation folds, the dataset had an average skewness of $2.53 (\pm 0.09)$ and kurtosis of $4.42 (\pm 0.46)$ before balancing. After applying SMOTE, the skewness increased to $2.98 (\pm 0.08)$ and kurtosis to $6.89 (\pm 0.48)$.

Table 1 presents the average performance of the nine algorithms on the unseen / testing data as reported by the nested cross-validation. The Naive Bayes (NB) model trained on the domain experts' features reported the overall best results, with Accuracy 0.93 (s.e. 0.08), Area Under the Receiver Operating Characteristics Curve (AUC): 0.96 (s.e. 0.04), Mathew's Correlation Coefficient (MCC) 0.86 (s.e. 0.16), F1-score 0.91 (s.e. 0.1), Negative Predictive value (NPV) 0.98 (s.e. 0.03), Positive Predictive value (PPV) 0.86 (s.e. 0.15), Specificity (SP) 0.90 (s.e. 0.12), and Sensitivity (SE) 0.97 (s.e. 0.06). Figure 2 illustrates the ROC curve of NB for each of the five cross-validation folds. Moreover, Fig. 3 shows the best-case confusion matrix for NB for each of the five cross-validation folds. In each matrix, the model achieves a high count of true-positives and true-negatives showing that the classifier consistently distinguishes MPS versus control with few misclassification. For the AdaBoost model, the highest AUC reported is 0.95 by training the model using Chi-Square and expert features. For the decision tree, KNN, and MLP, mutual information features provided the best results over all the evaluation metrics (decision tree: Accuracy: 0.87, AUC: 0.9, F1-score: 0.75, and MCC: 0.84; KNN Accuracy: 0.84, AUC: 0.9, F1-score: 0.7, and MCC: 0.81 and MLP: Accuracy: 0.93, AUC: 0.95, F1-score: 0.84, and MCC: 0.9). While for Gradient Boosting and Random Forest, the best-reported results were based on the models trained on the Chi-square features; where Gradient Boosting reported Accuracy: 0.87, AUC: 0.92, F1-score: 0.74, and MCC: 0.84 and Random Forest reported Accuracy: 0.85, AUC: 0.91, F1-score: 0.72, and MCC: 0.83. Finally, for SVC, select from a model based on logistic regression features selection stated the highest performance reported for detecting MPS patients: Accuracy: 0.86, AUC: 0.92, F1-score: 0.75, and MCC: 0.84.

After finding the best model based on the best-reported performance in the evaluation metrics, which is NB trained on the domain expert, we conduct further analysis to understand and interpret the models' decisions and explain why the model reached the reported conclusions using the SHAP and LIME summary plots. Figure 4 shows the importance of the features in order from highest to lowest reported by SHAP analysis. Following is the order of top 15 features from the most important to the least: acute gingivitis, plaque-induced, accretions on teeth, body mass index (BMI) pediatric, greater than or equal to 95th percentile for age, chronic gingivitis, plaque-induced, dental caries on smooth surface penetrating into dentin, acute pharyngitis, unspecified, acute tonsillitis due to other specified organisms, dental caries extending into dentine, acute tonsillitis due to other specified organisms, acute tonsillitis, unspecified, acute pharyngitis, acute bronchitis, nasal congestion, chronic rhinitis, and wheezing. Additionally, the feature importance of the best performing model is visualized using LIME method in Fig. 5. The top 15 features ordered by LIME from the most important to the least are: accretions on teeth, acute pharyngitis, unspecified, acute gingivitis, plaque-induced, acute pharyngitis, chronic gingivitis, plaque-induced, dental caries extending into dentine, dental caries on smooth surface penetrating into dentin, body mass index (BMI) pediatric, greater than or equal to 95th percentile for age, nasal congestion, acute tonsillitis due to other specified organisms, acute bronchitis, chronic rhinitis, acute tonsillitis due to other specified organisms, acute upper respiratory infection, and acute tonsillitis.

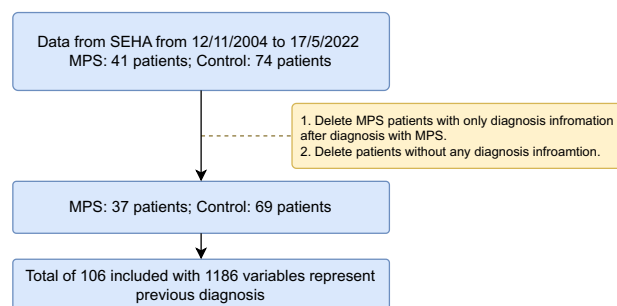


Fig. 1. Study cohort selection flow diagram. Patient selection criteria and training and testing sets splitting of the dataset.

Models	Metrics	All	Chi-Square	Experts	Select From Model	Lasso	Mutual Info	BAT	GA
AdaBoost	Accuracy	0.85 (0.13)	0.89 (0.09)	0.89 (0.10)	0.86 (0.14)	0.86 (0.14)	0.88 (0.08)	0.85 (0.07)	0.84 (0.11)
	AUC	0.91 (0.10)	0.95 (0.06)	0.95 (0.04)	0.92 (0.10)	0.89 (0.12)	0.95 (0.06)	0.91 (0.05)	0.91 (0.09)
	Matthews	0.71 (0.25)	0.77 (0.19)	0.77 (0.20)	0.73 (0.26)	0.72 (0.29)	0.75 (0.16)	0.71 (0.15)	0.69 (0.20)
	F1-score	0.82 (0.15)	0.85 (0.12)	0.86 (0.12)	0.83 (0.15)	0.83 (0.17)	0.84 (0.10)	0.81 (0.09)	0.81 (0.12)
	NPV	0.94 (0.09)	0.94 (0.06)	0.95 (0.07)	0.94 (0.09)	0.93 (0.11)	0.95 (0.07)	0.95 (0.07)	0.93 (0.04)
	PPV	0.75 (0.18)	0.81 (0.13)	0.81 (0.14)	0.77 (0.18)	0.77 (0.19)	0.78 (0.09)	0.73 (0.10)	0.75 (0.17)
	SP	0.81 (0.16)	0.88 (0.08)	0.87 (0.11)	0.83 (0.16)	0.84 (0.14)	0.85 (0.08)	0.81 (0.08)	0.81 (0.16)
	SE	0.92 (0.11)	0.90 (0.11)	0.92 (0.11)	0.92 (0.11)	0.90 (0.16)	0.92 (0.11)	0.92 (0.11)	0.89 (0.06)
Decision Tree	Accuracy	0.84 (0.11)	0.85 (0.13)	0.83 (0.11)	0.86 (0.12)	0.84 (0.11)	0.87 (0.12)	0.84 (0.18)	0.83 (0.11)
	AUC	0.87 (0.11)	0.88 (0.13)	0.88 (0.10)	0.88 (0.11)	0.89 (0.08)	0.90 (0.11)	0.84 (0.19)	0.83 (0.11)
	Matthews	0.70 (0.20)	0.70 (0.27)	0.63 (0.25)	0.74 (0.22)	0.71 (0.20)	0.75 (0.22)	0.68 (0.35)	0.66 (0.21)
	F1-score	0.80 (0.13)	0.81 (0.16)	0.75 (0.18)	0.84 (0.13)	0.81 (0.12)	0.84 (0.13)	0.80 (0.21)	0.76 (0.16)
	NPV	0.93 (0.08)	0.93 (0.11)	0.89 (0.11)	0.96 (0.05)	0.95 (0.07)	0.95 (0.07)	0.90 (0.14)	0.89 (0.08)
	PPV	0.79 (0.21)	0.75 (0.18)	0.74 (0.13)	0.77 (0.18)	0.75 (0.18)	0.79 (0.17)	0.78 (0.22)	0.81 (0.20)
	SP	0.82 (0.19)	0.83 (0.13)	0.85 (0.09)	0.81 (0.18)	0.80 (0.17)	0.84 (0.17)	0.84 (0.20)	0.85 (0.18)
	SE	0.86 (0.18)	0.90 (0.16)	0.78 (0.24)	0.95 (0.07)	0.92 (0.13)	0.92 (0.13)	0.84 (0.22)	0.78 (0.22)
NB	Accuracy	0.85 (0.10)	0.88 (0.08)	0.93 (0.08)	0.91 (0.09)	0.85 (0.12)	0.86 (0.06)	0.81 (0.11)	0.83 (0.13)
	AUC	0.89 (0.09)	0.92 (0.09)	0.96 (0.04)	0.92 (0.07)	0.92 (0.12)	0.93 (0.07)	0.80 (0.12)	0.84 (0.13)
	Matthews	0.71 (0.20)	0.75 (0.16)	0.86 (0.16)	0.80 (0.19)	0.69 (0.24)	0.73 (0.12)	0.60 (0.24)	0.66 (0.25)
	F1-score	0.81 (0.13)	0.84 (0.10)	0.91 (0.10)	0.88 (0.12)	0.81 (0.14)	0.83 (0.07)	0.73 (0.15)	0.79 (0.15)
	NPV	0.94 (0.06)	0.95 (0.07)	0.98 (0.03)	0.95 (0.07)	0.92 (0.08)	0.97 (0.05)	0.86 (0.10)	0.90 (0.09)
	PPV	0.76 (0.17)	0.78 (0.09)	0.86 (0.15)	0.84 (0.13)	0.76 (0.16)	0.74 (0.09)	0.75 (0.18)	0.75 (0.19)
	SP	0.83 (0.15)	0.85 (0.08)	0.90 (0.12)	0.90 (0.08)	0.84 (0.13)	0.81 (0.08)	0.85 (0.10)	0.83 (0.16)
	SE	0.89 (0.12)	0.92 (0.11)	0.97 (0.06)	0.92 (0.11)	0.87 (0.13)	0.95 (0.07)	0.74 (0.18)	0.84 (0.14)
Gradient Boosting	Accuracy	0.85 (0.14)	0.87 (0.11)	0.86 (0.12)	0.84 (0.14)	0.83 (0.12)	0.86 (0.12)	0.84 (0.12)	0.86 (0.14)
	AUC	0.89 (0.08)	0.92 (0.06)	0.92 (0.07)	0.90 (0.10)	0.89 (0.11)	0.89 (0.07)	0.91 (0.08)	0.90 (0.12)
	Matthews	0.70 (0.29)	0.74 (0.21)	0.72 (0.22)	0.67 (0.29)	0.66 (0.26)	0.72 (0.25)	0.67 (0.25)	0.72 (0.29)
	F1-score	0.81 (0.17)	0.84 (0.12)	0.82 (0.13)	0.79 (0.18)	0.78 (0.15)	0.82 (0.15)	0.79 (0.16)	0.82 (0.18)
	NPV	0.92 (0.10)	0.95 (0.08)	0.92 (0.08)	0.89 (0.11)	0.92 (0.11)	0.94 (0.09)	0.91 (0.10)	0.92 (0.10)
	PPV	0.76 (0.19)	0.77 (0.13)	0.79 (0.17)	0.77 (0.19)	0.72 (0.14)	0.77 (0.17)	0.74 (0.15)	0.79 (0.22)
	SP	0.84 (0.14)	0.84 (0.12)	0.86 (0.14)	0.85 (0.13)	0.81 (0.11)	0.84 (0.14)	0.84 (0.09)	0.85 (0.15)
	SE	0.87 (0.15)	0.92 (0.11)	0.87 (0.13)	0.82 (0.17)	0.87 (0.18)	0.89 (0.15)	0.84 (0.17)	0.87 (0.15)
K-Neighbors	Accuracy	0.77 (0.15)	0.83 (0.11)	0.81 (0.09)	0.78 (0.15)	0.81 (0.14)	0.84 (0.10)	0.73 (0.15)	0.67 (0.03)
	AUC	0.85 (0.11)	0.89 (0.10)	0.88 (0.07)	0.86 (0.11)	0.86 (0.13)	0.90 (0.09)	0.84 (0.14)	0.82 (0.08)
	Matthews	0.60 (0.22)	0.70 (0.18)	0.67 (0.14)	0.63 (0.23)	0.62 (0.28)	0.70 (0.17)	0.57 (0.23)	0.48 (0.05)
	F1-score	0.75 (0.13)	0.81 (0.11)	0.79 (0.09)	0.76 (0.15)	0.76 (0.17)	0.81 (0.10)	0.73 (0.14)	0.67 (0.03)
	NPV	0.96 (0.05)	0.96 (0.05)	0.98 (0.04)	0.97 (0.07)	0.90 (0.09)	0.96 (0.05)	0.97 (0.06)	0.98 (0.05)
	PPV	0.64 (0.17)	0.72 (0.15)	0.67 (0.12)	0.66 (0.20)	0.71 (0.19)	0.72 (0.12)	0.59 (0.18)	0.51 (0.02)
	SP	0.67 (0.24)	0.77 (0.16)	0.72 (0.13)	0.68 (0.24)	0.80 (0.15)	0.78 (0.13)	0.60 (0.22)	0.51 (0.06)
	SE	0.94 (0.08)	0.95 (0.07)	0.97 (0.06)	0.94 (0.13)	0.84 (0.15)	0.95 (0.07)	0.97 (0.06)	0.97 (0.06)
Logistic Regression	Accuracy	0.83 (0.13)	0.86 (0.11)	0.88 (0.08)	0.84 (0.11)	0.84 (0.12)	0.85 (0.10)	0.85 (0.13)	0.87 (0.14)
	AUC	0.93 (0.07)	0.92 (0.13)	0.95 (0.06)	0.93 (0.08)	0.91 (0.10)	0.93 (0.09)	0.90 (0.13)	0.92 (0.11)
	Matthews	0.69 (0.22)	0.72 (0.21)	0.78 (0.14)	0.70 (0.20)	0.69 (0.22)	0.72 (0.18)	0.71 (0.24)	0.75 (0.27)
	F1-score	0.81 (0.13)	0.83 (0.12)	0.85 (0.09)	0.81 (0.12)	0.81 (0.13)	0.82 (0.10)	0.82 (0.14)	0.84 (0.16)
	NPV	0.94 (0.06)	0.93 (0.07)	0.98 (0.03)	0.95 (0.05)	0.93 (0.07)	0.96 (0.05)	0.94 (0.09)	0.94 (0.09)
	PPV	0.73 (0.19)	0.78 (0.17)	0.77 (0.14)	0.74 (0.17)	0.75 (0.17)	0.73 (0.12)	0.75 (0.15)	0.78 (0.19)
	SP	0.78 (0.19)	0.84 (0.14)	0.83 (0.13)	0.80 (0.16)	0.81 (0.16)	0.80 (0.14)	0.81 (0.15)	0.84 (0.16)
	SE	0.92 (0.07)	0.89 (0.11)	0.97 (0.06)	0.92 (0.07)	0.89 (0.12)	0.95 (0.07)	0.92 (0.11)	0.92 (0.11)
Continued									

MLP	Accuracy	0.90 (0.13)	0.89 (0.10)	0.91 (0.10)	0.90 (0.11)	0.84 (0.10)	0.93 (0.06)	0.91 (0.09)	0.87 (0.13)
	AUC	0.95 (0.06)	0.95 (0.07)	0.94 (0.07)	0.95 (0.06)	0.92 (0.07)	0.95 (0.06)	0.95 (0.06)	0.93 (0.08)
	Matthews	0.81 (0.24)	0.78 (0.21)	0.81 (0.19)	0.80 (0.21)	0.67 (0.19)	0.84 (0.13)	0.82 (0.17)	0.74 (0.23)
	F1-score	0.88 (0.15)	0.86 (0.13)	0.88 (0.12)	0.88 (0.13)	0.80 (0.11)	0.90 (0.08)	0.88 (0.11)	0.84 (0.14)
	NPV	0.96 (0.05)	0.95 (0.07)	0.94 (0.06)	0.95 (0.05)	0.92 (0.06)	0.97 (0.04)	0.96 (0.06)	0.92 (0.05)
	PPV	0.84 (0.20)	0.81 (0.16)	0.87 (0.15)	0.84 (0.18)	0.74 (0.12)	0.86 (0.10)	0.85 (0.15)	0.83 (0.20)
	SP	0.87 (0.18)	0.87 (0.12)	0.91 (0.12)	0.88 (0.15)	0.83 (0.11)	0.91 (0.06)	0.90 (0.12)	0.87 (0.18)
	SE	0.95 (0.07)	0.92 (0.11)	0.90 (0.11)	0.92 (0.07)	0.87 (0.09)	0.95 (0.07)	0.92 (0.11)	0.86 (0.10)
Random Forest	Accuracy	0.85 (0.12)	0.85 (0.11)	0.82 (0.14)	0.85 (0.12)	0.84 (0.11)	0.83 (0.13)	0.87 (0.15)	0.83 (0.12)
	AUC	0.90 (0.11)	0.91 (0.09)	0.90 (0.08)	0.90 (0.10)	0.88 (0.10)	0.90 (0.10)	0.91 (0.10)	0.90 (0.10)
	Matthews	0.73 (0.21)	0.72 (0.20)	0.70 (0.19)	0.73 (0.21)	0.70 (0.20)	0.70 (0.22)	0.74 (0.30)	0.65 (0.23)
	F1-score	0.83 (0.12)	0.83 (0.12)	0.81 (0.11)	0.83 (0.12)	0.81 (0.12)	0.81 (0.13)	0.84 (0.18)	0.78 (0.15)
	NPV	0.96 (0.05)	0.96 (0.05)	0.98 (0.03)	0.96 (0.05)	0.95 (0.05)	0.96 (0.06)	0.93 (0.11)	0.89 (0.07)
	PPV	0.75 (0.17)	0.74 (0.17)	0.71 (0.16)	0.75 (0.17)	0.74 (0.17)	0.72 (0.18)	0.80 (0.21)	0.76 (0.18)
	SP	0.80 (0.17)	0.80 (0.16)	0.74 (0.23)	0.80 (0.17)	0.80 (0.17)	0.77 (0.19)	0.85 (0.15)	0.84 (0.16)
	SE	0.95 (0.07)	0.95 (0.07)	0.97 (0.06)	0.95 (0.07)	0.92 (0.08)	0.95 (0.07)	0.90 (0.16)	0.81 (0.16)
SVC	Accuracy	0.83 (0.11)	0.85 (0.13)	0.85 (0.10)	0.86 (0.10)	0.85 (0.11)	0.86 (0.10)	0.79 (0.13)	0.81 (0.18)
	AUC	0.93 (0.05)	0.92 (0.09)	0.92 (0.09)	0.92 (0.09)	0.91 (0.10)	0.90 (0.13)	0.85 (0.13)	0.88 (0.15)
	Matthews	0.70 (0.18)	0.71 (0.25)	0.71 (0.20)	0.75 (0.16)	0.72 (0.20)	0.72 (0.22)	0.57 (0.27)	0.62 (0.35)
	F1-score	0.81 (0.11)	0.82 (0.15)	0.82 (0.12)	0.84 (0.10)	0.83 (0.12)	0.82 (0.14)	0.73 (0.16)	0.77 (0.21)
	NPV	0.97 (0.05)	0.94 (0.09)	0.95 (0.08)	0.98 (0.03)	0.96 (0.05)	0.94 (0.09)	0.87 (0.11)	0.88 (0.13)
	PPV	0.72 (0.17)	0.75 (0.18)	0.73 (0.12)	0.75 (0.16)	0.74 (0.17)	0.76 (0.15)	0.69 (0.15)	0.73 (0.22)
	SP	0.77 (0.18)	0.81 (0.16)	0.81 (0.11)	0.80 (0.16)	0.80 (0.16)	0.84 (0.11)	0.80 (0.14)	0.81 (0.19)
	SE	0.95 (0.07)	0.92 (0.11)	0.92 (0.11)	0.97 (0.06)	0.95 (0.07)	0.90 (0.16)	0.79 (0.18)	0.81 (0.20)

Table 1. The performance of different machine learning models for MPS prediction in the testing set.

Discussion

Mucopolysaccharidosis (MPS) represents a group of rare inherited metabolic disorders characterized by the deficiency of lysosomal enzymes essential for the degradation of glycosaminoglycans (GAG), leading to their accumulation within cells and subsequent systemic symptoms^{26,27}. MPS encompasses seven subtypes, each associated with distinct enzyme deficiencies and clinical manifestations²⁸. The prevalence of MPS varies significantly across different populations, with certain ethnicities exhibiting higher incidences²⁹. The infrequency of MPS poses significant challenges in diagnosis and management, further complicated by the wide spectrum of clinical presentations.

To the best of our knowledge, this is the first study that trained and compared the performance of different machine learning models to predict MPS cases. We trained nine machine learning models, namely, AdaBoost, Decision Tree, Gaussian Naive Bayes, Gradient Boosting Classifier, K-nearest Neighbors' Algorithm, Logistic Regression, Multi-layer Perceptron Classifier, Random Forests, and Support Vector Classification in a combination of five features selection methods: Chi-square feature selection, domain experts feature set, select from the model (logistic Regression), Lasso 5-fold feature selection, Mutual information, Bat Algorithm, Genetic Algorithm using patients past medical history from SEHA medical records, UAE. The models were compared using various evaluation metrics such as accuracy, Area Under the Receiver Operating Characteristics Curve, F1-score, Matthews Correlation Coefficient, Negative Predictive Value, Positive Predictive Value, Specificity and Sensitivity on the unseen datasets.

Overall, NB trained using domain expert features reported the highest performance Accuracy 0.93 (s.e. 0.08), AUC 0.96 (s.e. 0.04), Mathew's coefficient 0.86 (s.e. 0.16), F1-score 0.91 (s.e. 0.1), NPV 0.98 (s.e. 0.03), PPV 0.86 (s.e. 0.15), SP 0.90 (s.e. 0.12), and SE 0.97 (s.e. 0.06). Figures 4 and 5 represent the features selected by the NB model to detect MPS patients at early stages based on expert features. Both SHAP and LIME identified a highly consistent set of key features. The highest selected feature, interestingly, were the dental manifestations of the disease, mainly acute gingivitis, accretions on teeth, chronic gingivitis, and dental caries. It is well known that patients with MPS present with dental anomalies, deviations in eruption, malocclusions, TMJ pathology, macroglossia, gingival hyperplasia, and increased risk of caries and periodontal disease³⁰. One of the studies reported 76% of patients with MPS IV had experienced dental caries and all patients with MPS showed evidence of a generalized unspecified enamel defect, and 43% of them exhibited marginal gingivitis³¹.

Patients with MPS have poor oral hygiene. These findings maybe as result of difficulties in maintaining oral hygiene since some of these patients have intellectual disabilities or restriction of joint movement which might affect their brushing techniques, poor follow up with dentist since they do have multiple significant comorbidities. Some of the dental procedure require general anesthesia which is challenging in those patients because of airway involvement of the disease. Most of the time given the complexity of the disease, their follow up is also limited to tertiary centers^{31–34}.

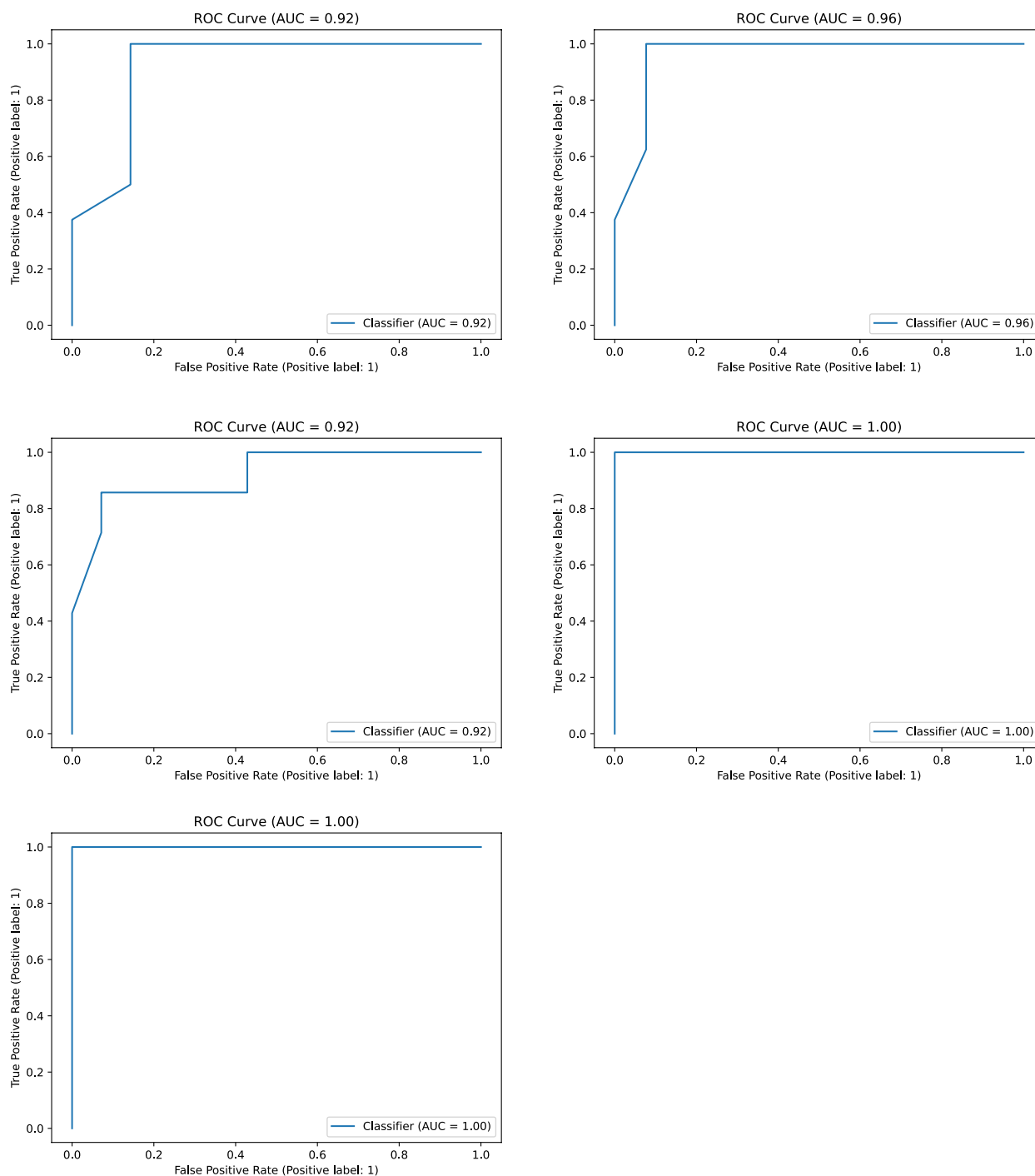


Fig. 2. AUC curve of the best performing model across five folds.

The other identified features of the disease were acute pharyngitis, acute tonsillitis, acute bronchitis, nasal congestion. Indeed, recurrent respiratory infection is the most common feature of these disorders. These features have been reported in most of the studies, it considered one of the early presentation of patients with MPS diseases in the first two year of life^{35–37}.

Body mass index pediatric, greater or equal to the 95 percentiles, means that the cases are within the obese range of BMI. Patel et al studied the growth parameter of patients with MPS II and compared it with normal control. They have noted that 97% of studied patients had a BMI higher than the mean BMI of the normal control³⁸. Another study investigated the natural history of growth parameter from untreated males followed prospectively in the Hunter Outcome Survey registry and found that BMI was above average throughout childhood until approximately 14–16 years of age³⁹.

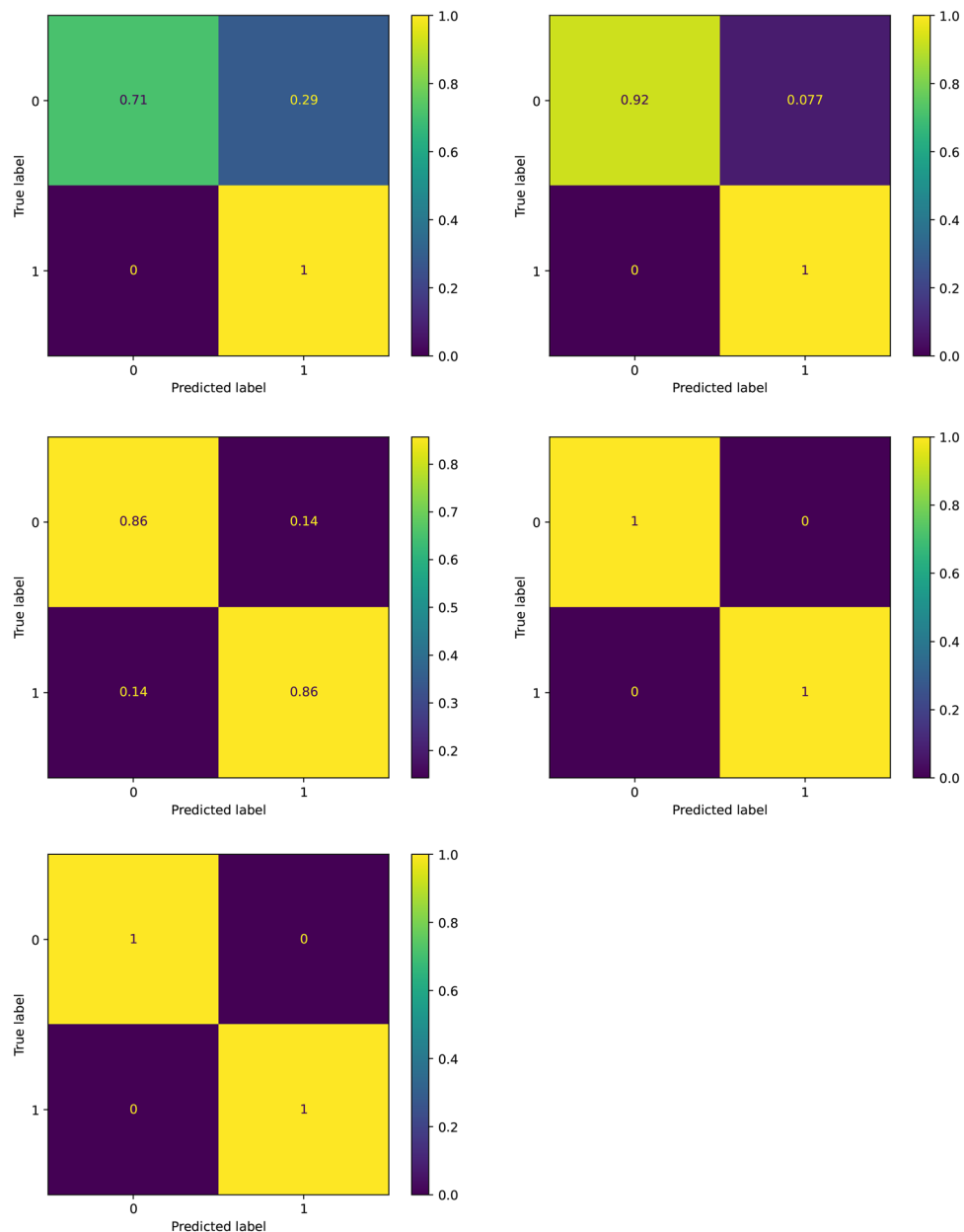


Fig. 3. Confusion matrix of the best performing model across five folds.

Most of these features are non-specific to MPS but the combination of these symptoms yields better prediction of MPS in our study. It also reflects the local practice of clinicians and their clinical documentation. Moreover, it highlights the significant burden of dental manifestations in our patients. In comparison, the data analysis of Hurler registry showed that umbilical and inguinal hernia as well as coarse facies and corneal clouding are among the early symptoms of patients with MPS I⁴⁰.

On the other hand, based on the data from the registry of Hunter Outcome Survey the following symptoms were considered helpful in the diagnosis of the disease: facial dysmorphism, nasal obstruction or rhinorrhea, enlarged tongue, enlarged liver, enlarged spleen, joint stiffness which were given the weight of 2 while the other symptoms of hernia, hearing impairment, enlarged tonsils, airway obstruction or sleep apnea were given weight of 1. A mnemonic screening tool was developed based on these data with total score of 6 or greater with high risk of the MPS II⁴¹. Our model identified nasal congestion one of the high risk feature of the disease.

This research offers a cost-effective screening method for RD participants. It utilizes the current medical record system powered by AI models, eliminating the need for clinical experts to manually identify and flag suspicious undiagnosed cases. We validated the applicability of machine learning models for predicting MPS cases using patients' past medical history from SEHA electronic health records, United Arab Emirates cohort. Our finding confirms the power of machine learning to detect rare disease cases, as reported by different evaluation metrics that are used to compare different ML models' performance with unseen data.

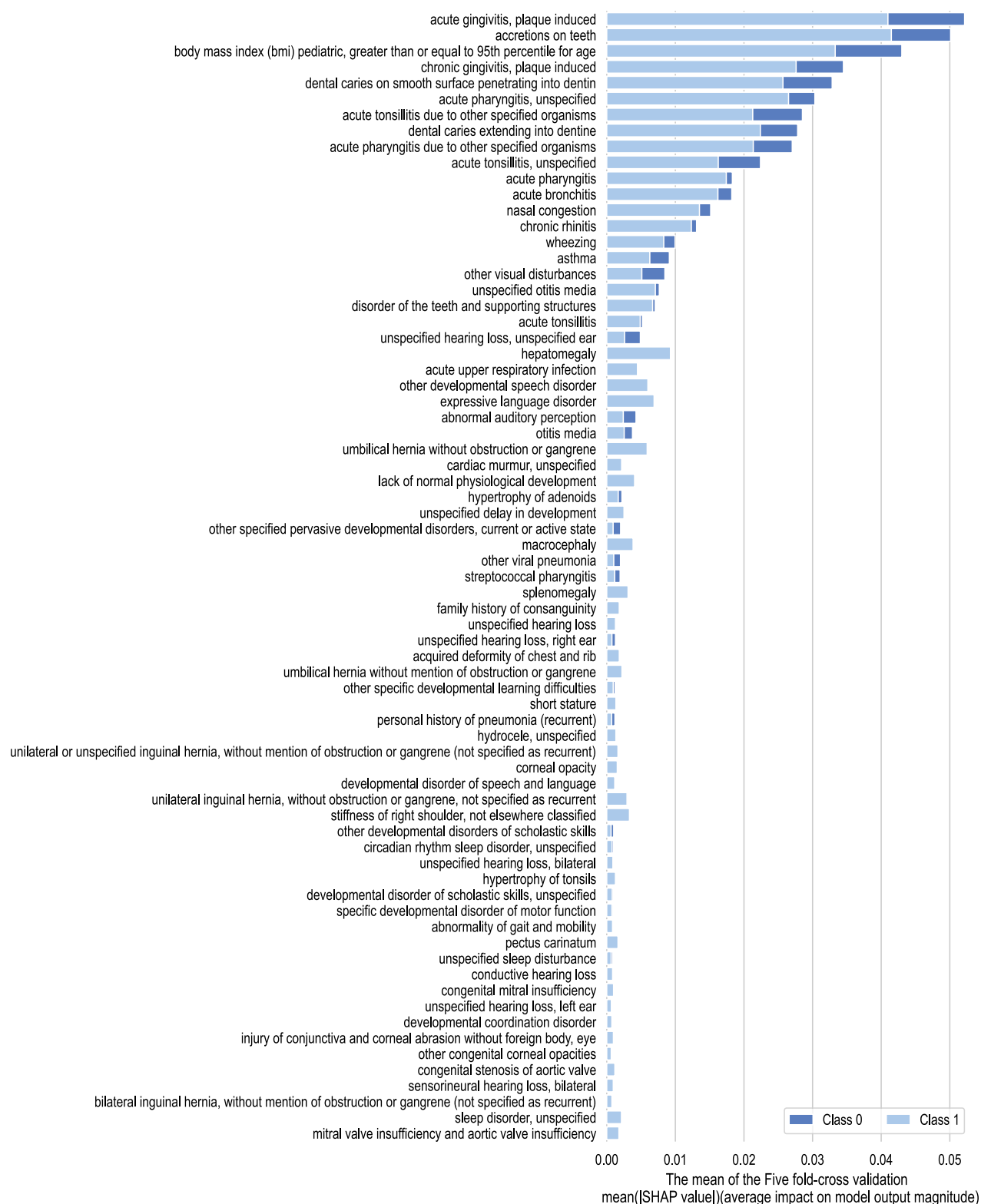


Fig. 4. Variables Importance: Variable importance plot for Naive Bayes trained on the features extracted from a domain expert using SHAP analysis.

Despite these strengths, our study has several limitations that could be improved upon in future studies. First, our study relies exclusively on diagnostic codes from a single healthcare system (SEHA) which is limited and may not capture relevant clinical differences. Additionally, because there are no prior studies on using machine learning to diagnose MPS using historical EHR data, we lacked an established benchmark for

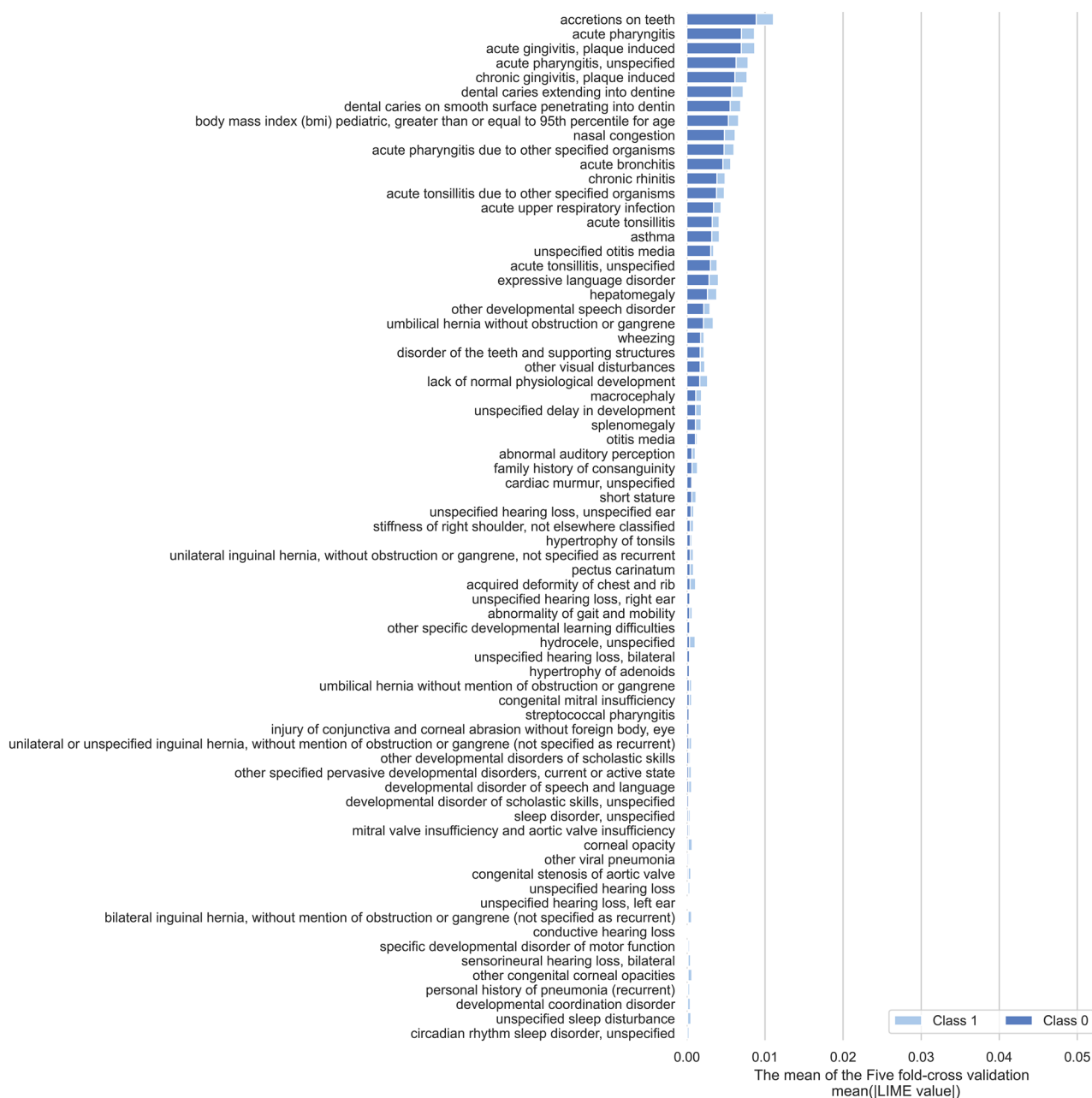


Fig. 5. Variables Importance: Variable importance plot for Naive Bayes trained on the features extracted from a domain expert using LIME Analysis.

comparison and validation. Moreover, the study did not incorporate demographic information or any other data modalities which might improve the model's performance and report more significant predictors for diagnosing MPS patients. The study also does not distinguish between MPS subtypes, which ultimately subtype-specific outcomes or predictive features. Furthermore, the study's heavy reliance on EHR, which tends to naturally have noise, missing data, could have affected the feature selection and model accuracy. Finally, external validation on multi-center, ethnically diverse cohorts is necessary to confirm generalizability before using this framework in border clinical settings.

Conclusion

In conclusion, this study presented a machine learning framework for the early diagnosis of MPS relying on patients' historical medical diagnoses extracted from EHR data. We evaluated multiple ML models in combination with different feature selection algorithms to efficiently diagnose patients. Our results demonstrate that incorporating domain expert-selected features with a Naïve Bayes model achieved the highest diagnostic accuracy in identifying MPS patients. Additionally, the feature importance of the best-performing model supported the common clinical manifestation presented in MPS disease, highlighting the model's capabilities in

capturing MPS disease pathology. The obtained results demonstrate the potential of utilizing ML models with historical diagnostic data in RD diagnosis, particularly MPS disease, enabling more efficient and cost-effective screening tools. Future research could involve integrating additional clinical data, such as laboratory results, imaging, and genetic information, with diagnostic features to further enhance predictive performance and contextual insight. Moreover, extending the framework to include larger, multi-center datasets could improve generalizability, while exploring MPS subtype classification could offer more precise and personalized diagnostic support.

Methods

Data source and study cohort

In this retrospective study, we extracted anonymized patients' medical records from the Abu Dhabi Health Services Company (SEHA) healthcare system. The SEHA dataset is a high-dimensional UAE population healthcare data source that includes rich patients' medical information such as demographics, comorbidity, symptoms upon admission, growth parameters, laboratory results, and medications. The final dataset included 106 patients, of which 37 were diagnosed with MPS and 69 were controls.

Study variables

The study outcome was the diagnosis of MPS; the variable was dichotomized (0 and 1), where one indicates MPS positive and zero otherwise. The covariates or independent variables included all patients' medical history, which was also dichotomized to indicate the disease's presence or absence. In total, we extracted 1186 covariates/comorbidities covering a wide range of medical conditions. Then after feature selection, the selected features were used to train the nine machine learning models. Table 2 presents the features identified by the domain expert.

Machine learning framework

Figure 6 presents the study pipeline. We used Scikit-learn v1.3.0 of Python programming language v3.9.17 to implement the machine learning models along with Hyperopt v0.2.7 and Imbalanced-learn v0.11.0 python packages. We utilized nested-cross validation (double cross-validation) to stratify the dataset into training and testing sets, as well as to optimize feature selection algorithms and machine learning model parameters. The nested-cross validation consists of an outer loop and an inner loop. The outer loop is used to estimate

List of clinical covariates			
Unspecified hearing loss	Sleep disorder, unspecified	Stiffness of right shoulder, not elsewhere classified	Other visual disturbances
Unspecified sleep disturbance	Pectus carinatum	Unspecified hearing loss, left ear	Splenomegaly
Developmental disorder of speech and language	Chronic rhinitis	Hypertrophy of tonsils	Acute tonsillitis due to other specified organisms
Acute bronchitis	Unilateral inguinal hernia, without obstruction or gangrene, not specified as recurrent	Other developmental disorders of scholastic skills	Asthma
Circadian rhythm sleep disorder, unspecified	Unspecified hearing loss, unspecified ear	Hepatomegaly	Unilateral or unspecified inguinal hernia, without mention of obstruction or gangrene (not specified as recurrent)
Cardiac murmur, unspecified	Specific developmental disorder of motor function	Abnormality of gait and mobility	Other specific developmental learning difficulties
Nasal congestion	Disorder of the teeth and supporting structures	Acute tonsillitis, unspecified	Accretions on teeth
Dental caries extending into dentine	Macrocephaly	Wheezing	Unspecified otitis media
Bilateral inguinal hernia, without mention of obstruction or gangrene (not specified as recurrent)	Streptococcal pharyngitis	Acute pharyngitis	Other congenital corneal opacities
Other viral pneumonia	Umbilical hernia without obstruction or gangrene	Congenital mitral insufficiency	Personal history of pneumonia (recurrent)
Acute pharyngitis due to other specified organisms	Corneal opacity	Other specified pervasive developmental disorders, current or active state	Otitis media
Unspecified hearing loss, right ear	Lack of normal physiological development	Hydrocele, unspecified	Developmental disorder of scholastic skills, unspecified
Short stature	Conductive hearing loss	Acute upper respiratory infection	Unspecified hearing loss, bilateral
Expressive language disorder	Mitral valve insufficiency and aortic valve insufficiency	Umbilical hernia without mention of obstruction or gangrene	Hypertrophy of adenoids
Sensorineural hearing loss, bilateral	Acquired deformity of chest and rib	Acute tonsillitis	Developmental coordination disorder
Other developmental speech disorder	Family history of consanguinity	Abnormal auditory perception	Acute gingivitis, plaque induced
Body mass index (bmi) pediatric, greater than or equal to 95th percentile for age	Chronic gingivitis, plaque induced	Dental caries on smooth surface penetrating into dentin	Injury of conjunctiva and corneal abrasion without foreign body, eye
Unspecified delay in development	Congenital stenosis of aortic valve	Acute pharyngitis, unspecified	

Table 2. Clinical covariates selected based on domain expert knowledge.

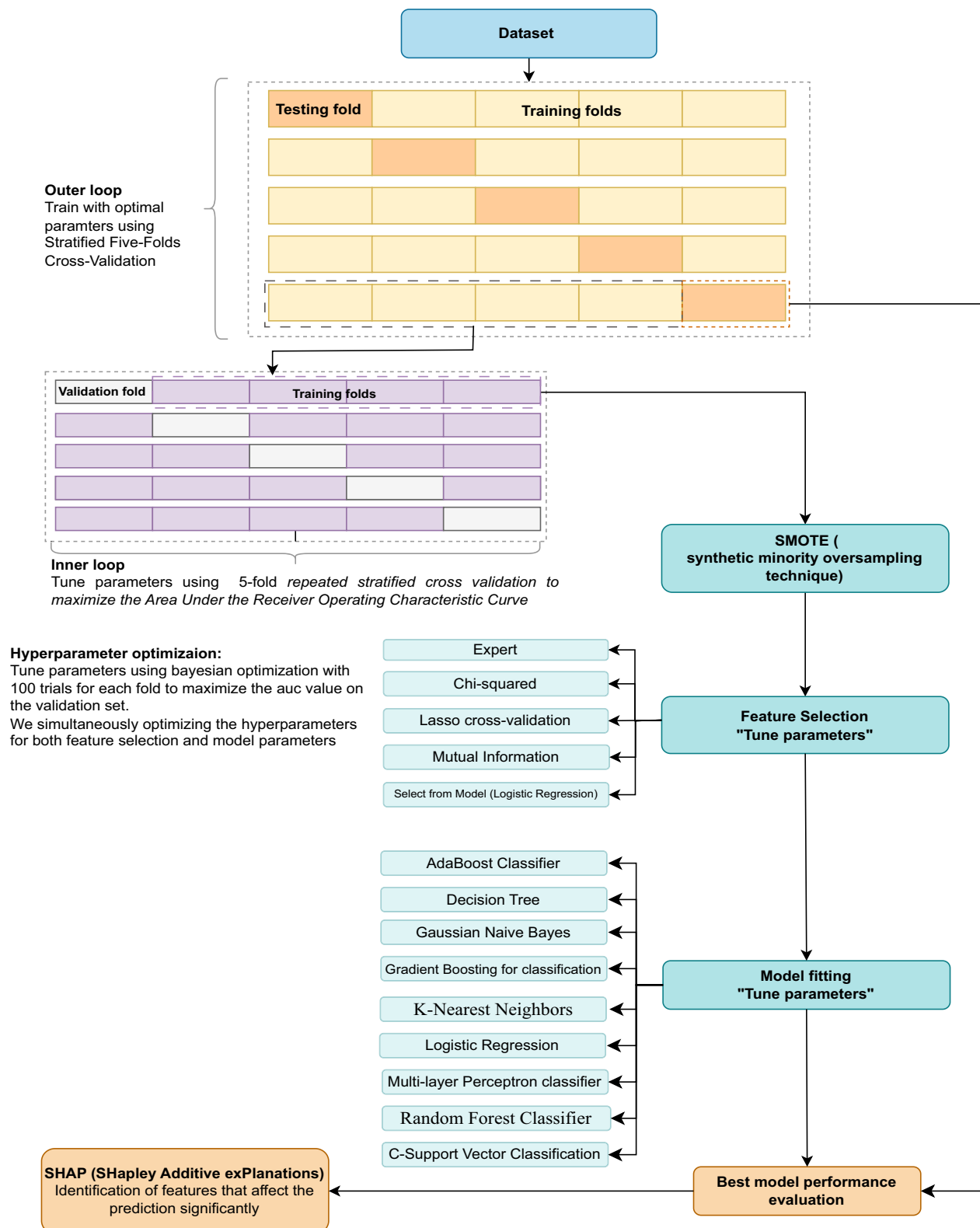


Fig. 6. Machine learning workflow.

the unbiased prediction accuracy of the machine learning models⁴²; in this stage, we utilize stratified five-fold cross-validation to create training folds that the machine learning models use to learn the new representation that distinguishes MPS patients from the control patients. After training the models, we evaluated the trained models' performance on the unseen testing fold. Based on the performance of different models in the testing set, we compared them to select the best model to tackle this challenge. Stratification sampling was selected for

cross-validation to guarantee the training and testing are representative of the different groups' distribution in our cohorts⁴³. The inner loop of the nested-cross validation is responsible for hyperparameter optimization for both feature selection algorithm parameters and machine learning model parameters⁴² – hyperparameters (Table 3 Models' hyperparameters). We applied Bayesian Optimization (BO)⁴⁴ with five-fold cross-validation. The hyperparameter-tuning process supports us in automatically selecting the set of parameters that gives almost optimal results using the training set. During this process, 100 sub-models were trained for each model and a feature selection-specific set of parameter values was selected during the optimization process. Since it is an optimization process, we set the objective function of the process to minimize the error on the validation set based on the Area Under the Receiver Operating Characteristic Curve (ROC AUC) score.

Imbalanced data

MPS is a rare condition; therefore, our training set is expected to be severely imbalanced in the number of samples between the two groups (a minority of the samples are MPS patients' records). It is well known that class imbalance affects machine learning techniques' decisions and directs the decision to the majority class⁴⁵. Therefore, several approaches were implemented to solve this problem, such as up-sampling the minority class or down-sampling the majority class. In this study, we applied the Synthetic Minority Oversampling Technique (SMOTE)⁴⁶ to increase the number of samples in the minority class. SMOTE works by generating new synthetic data points by linear interpolation of MPS records and K-Nearest Neighbors; for this work, we fixed K to five neighborhoods of samples to be used for generating the new synthetic samples.

Feature selection algorithms

Before training the model, we applied different supervised feature selection techniques to reduce the number of covariates for model training. The main objective of feature selection is to reduce the number of variables used to learn the new representation from the original dataset. This process helps to select the most informative features, exclude noisy or irrelevant features, prevent model overfitting, improve model performance, and minimize the computation power needed to run the code^{24,47,48}. We employed four automated feature selection methods and one domain expert-reported feature set. The automated feature selection techniques are Chi-Square feature selection, Lasso (Least Absolute Shrinkage and Selection Operator), mutual information (MI), and select from the model (logistic Regression).

- *Chi-square feature selection* is a univariate feature selection method that individually tests each feature and its relevance to output. A feature with a large Chi-Square value indicates a more important feature. The feature extracted is based on the hypothesis tests by selecting the statistically significant features where the significant level is set to p-value < 0.05^{24,48}. The formula for Chi-Square is:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where: O is the observed frequency and E is the expected frequency.

- *Lasso with 5-fold cross-validation* is an embedded-based method for feature selection. This method shrinks the regression coefficients to zero with respect to their contribution to the model output. Therefore, the algorithm selects the features based on their coefficient magnitude^{24,48}, due to the L1 regularization. The lasso loss function is calculated as follows:

$$\text{Minimize: } \sum_{i=1}^n \left(Y_i - \sum_{j=1}^p X_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (2)$$

Where n is the total number of samples, Y is the outcome, X is the independents/features, p is the total number of features, and β is the regression coefficient.

- *Mutual information* is a multivariate features selection method based on selecting the subset of features based on their inter-dependencies and the outcome. The method chooses features with the highest entropy-based estimation with the target⁴⁸. The MI represents as follows:

$$I(X;Y) = H(X) - H(X|Y) \quad (3)$$

Where $I(X; Y)$ is the MI for X, Y; $H(X)$ represents the entropy for X, while $H(X|Y)$ is the conditional entropy for X given Y.

- *Select from the model (logistic Regression)* is a wrapper-based method for feature selection; the algorithm selects the features based on the magnitude of their coefficients reported by logistic Regression⁴⁹. The logistic regression equation is:

Machine Learning Search Space		
Model	Parameter	Parameter Space
AdaBoost Classifier	Maximum number of estimators	{10, 50, 300, 750, 1200}
	Learning rate	{0.0001, 0.001, 0.01, 0.1, 1}
Decision Tree	Criterion	{Entropy, gini}
	Max depth	A numerical value like round (uniform(10, 1200)/10) * 10
	Max features	{Auto, sqrt, log2, None}
	Min samples leaf	Numerical space ranging from 0 to 0.5
	Min samples split	Numerical space ranging from 0 to 1
Gaussian Naive Bayes	Portion of the largest variance of all features that is added to variances for calculation stability	Numerical space ranging from 1e-09 to 1
Gradient Boosting Classifier	Learning rate	Numerical space ranging from 0.0001 to 1
	Number of estimators	{10, 50, 300, 750, 1200}
	Min samples split	Numerical space ranging from 0 to 1
	Min samples leaf	Numerical space ranging from 0 to 0.5
	Max depth	A numerical value like round(uniform(10, 1200)/10) * 10
k-nearest neighbors	Number of neighbors	{2,3,4,5,6,7,8,9,10}
	Weights	{Uniform, distance}
	Algorithm	{Auto, ball tree, kd tree, brute}
Logistic Regression	Penalty	{l1, l2, elasticnet, none}
Multi-layer Perceptron	Hidden layer sizes	{(100,), (100,100), (50,100), (50,50), (50,100,100), (50,50,50)}
	Alpha	Numerical space ranging from 0.0001 to 1
	Activation	{Relu, logistic, tanh}
	Solver	{Adam, sgd, lbfgs}
Random Forest Classifier	Criterion	{Entropy, gini}
	Max depth	A numerical value like round(uniform(10, 1200)/10) * 10
	Max features	{Auto, sqrt, log2, None}
	Min samples leaf	Numerical space ranging from 0 to 0.5
	Min samples split	Numerical space ranging from 0 to 1
	Number of estimators	{10, 50, 300, 750, 1200}
C-Support Vector Classifier	Regularization parameter	Numerical space ranging from 0.01 to 100
	Kernel	{Rbf, linear, sigmoid}
	Degree	Numerical value like round(uniform(2, 5))
	Gamma	{Scale, auto}
Feature Selection Search Space		
Chi-Square	K:Select features according to the k highest scores	{5, 10, 20, 50, 70}
Lasso	Alpha	Numerical space ranging from 0.0001 to 0.001
	Threshold	Numerical space ranging from 0.001 to 0.15
Select From Model (Logistic Regression)	C: Inverse of regularization strength	Numerical space ranging from 0.1 to 0.9
Mutual Information	K :Select features according to the k highest scores	{5, 10, 20, 50, 70}
Genetic Algorithm (GA)	Number of generations	Numerical space ranging from 20 to 40
	Number of individuals	Numerical space ranging from 10 to 40
	Parent selection method	{Roulette,tournament}
	Crossover probability	Numerical space ranging from 0.0 to 1.0
	Mutation probability	Numerical space ranging from 0.0 to 1.0
Bat Algorithm	Number of generations	Numerical space ranging from 20 to 40
	Number of bats	Numerical space ranging from 10 to 40
	Minimum frequency	Numerical space ranging from 0 to 2
	Maximum frequency	Numerical space ranging from 0 to 4
	Initial loudness of bats	Numerical space ranging from 0.0 to 1.0
	Initial pulse emission rate	Numerical space ranging from 0.0 to 1.0
	Loudness damping factor	Numerical space ranging from 0.0 to 1.0
	Pulse rate increasing factor	Numerical space ranging from 0.0 to 1.0

Table 3. Models' hyperparameters: machine learning models parameterized using a random search optimization algorithm of 100 different parameter settings with a 5-fold repeated stratified cross-validation to maximize the Area Under the Receiver Operating Characteristic Curve (AUC).

$$Y = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (4)$$

Where y is the outcome, X is the independent/feature variable and β is the logistic regression coefficient.

- *Bat Algorithm (BA)* is a metaheuristic inspired by bat echolocation. Each “bat” i has a position x_i , a velocity v_i , and a frequency f_i ⁵⁰. At iteration t , the velocity and position update is:

$$\begin{aligned} v_i^t &= v_i^{t-1} + (x_i^{t-1} - x_*^{t-1}) f_i^t, \\ x_i^t &= x_i^{t-1} + v_i^t. \end{aligned} \quad (5)$$

Where x_*^{t-1} is the best solution found up to iteration $t - 1$, and f_i^t is the bat’s frequency at iteration t .

- *Generic Algorithm (GA)* is an evolutionary method that evolves a population of solutions via selection, crossover, and mutation⁵¹. The probability of selecting individual i (with fitness $F(x_i)$) for reproduction is:

$$p_i = \frac{F(x_i)}{\sum_{j=1}^N F(x_j)}. \quad (6)$$

Where N is the population size, and p_i is the chance that x_i is chosen as a parent.

Machine learning algorithms

In this study, we developed and trained nine well-known machine learning models: the Adaptive Boost Classifier (AdaBoost), decision tree (DT), Naive Bayes (NB), gradient boosting classifier (XGBoost), k-nearest neighbors’ algorithm (KNN), logistic regression (LR), multi-layer Perceptron classifier (MLP), random forests (RF), and support vector classification (SVM). Following is a description of each method:

- *AdaBoost* is an ensemble machine learning model based on training multiple classifiers to improve model performance by learning from their errors in sequential matter. The AdaBoost algorithm was first proposed by Yoav Freund and Robert Shapire in 1995 stemming from an example of optimizing decisions of a horse-race gambler⁵². It combines the power of weak learning (a decision tree with a single level as a base classifier) to build a powerful and robust classifier with the iterative approach⁵³. AdaBoost has the power to adapt by improving the efficiency of classifiers such as decision trees but is very sensitive to noisy data and outliers. All weights start equally then on each round it increases the weights of misclassified samples, forcing the weak learner to focus on harder examples in subsequent iterations. The final model combines all weak learners aiming to minimize classification error⁵². The boosting algorithm is described as follows: Initialize the weight of distribution D on training example i on round $t = 1$:

$$D_{t=1}(i) = \frac{1}{m}. \quad (7)$$

For $t = 1, \dots, T$:

1. Train weak learner using distribution D_t .
2. Get weak hypothesis $h_t : X \rightarrow \{-1, +1\}$ with error ε_t :

$$\varepsilon_t = \Pr_{x \sim D_t} [h_t(x) \neq y]. \quad (8)$$

3. Choose:

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1 - \varepsilon_t}{\varepsilon_t} \right). \quad (9)$$

Update:

$$D_{t+1}(i) = \frac{D_t(i) e^{-\alpha_t y_i h_t(x_i)}}{Z_t}, \quad (10)$$

where Z_t is a normalization factor.

Output the final hypothesis:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right). \quad (11)$$

- *DT* is a non-parametric flowchart-like tree structure consisting of nodes (features), leaf nodes (outcomes), and a set of decision rules. They were first introduced in the 1960s and have been used in various disciplines because they are robust, easy to use, and free of ambiguity. Several statistical algorithms are used to build decision trees such as CART (Classification and Regression Trees), C4.5, CHAID (Chi-Squared Automatic Interaction Detection) and QUEST (Quick, Unbiased, Efficient, Statistical Tree). Given their advantages decision trees can have some limitations such as overfitting and underfitting when using a small dataset, and using correlated input variables may lead to misleading model improvements⁵⁴. The root node represents the top node in the tree; the tree is created in a recursive manner in which the rules are learned based on the values provided during the training time. Splitting refers to separating a single node (parent node) into many (child nodes) using input variables related to target variables by first identifying the most important input variables⁵⁴. The popular splitting rules are Gini impurity (Gini) and information gain (entropy)^{49,53}, which are expressed mathematically as follows:

$$\text{Gini}(E) = 1 - \sum_{i=1}^c p_i^2 \quad (12)$$

Where p is the probability that a sample belongs to a specific class (c).

$$\text{entropy} : H(X) = - \sum_{i=1}^n p(x_i) \log_2 p(x_i) \quad (13)$$

Where p is the probability of the entropy.

- *NB* is a simple supervised statistical model based on the Bayes theorem. The model is built on the assumption that the features are independent, in which each feature's effect is not related to/correlated with the other features^{49,53}. It works by calculating prior probabilities of a given class label and its likelihood probability and returns the conditional probability of a given target. NB is one of the simplest algorithms and is much faster than other supervised algorithms as it only calculates probabilities. However, it has limitations, such as the assumption that all features are independent, which may not hold in real-world data. Additionally, it assigns zero probability to categorical variables not seen during training, leading to an inability to predict such cases⁵⁵. The following is the mathematical formula for the NB model.

$$P(c | x) = \frac{P(x | c)P(c)}{P(x)} \quad (14)$$

Where $P(c | x)$ is the posterior probability of class c given x , $P(c)$ is the prior probability of class c , $P(x | c)$ is the probability of the x given c (likelihood), and $P(x)$ is the prior probability of the x .

- *XGBoost* (eXtreme Gradient Boosting) is an ensemble machine learning model where the models are trained sequentially. It was first introduced by Tianqi Chen in 2014 as part of the Distributed (Deep) Machine Learning Community (DMLC) group⁵⁶. XGBoost is particularly known for its high speed, performance and efficiency as it can utilize multiple CPU cores and it supports multiple loss functions making it adaptable. The prediction of XGBoost is based on the sum of outputs from multiple trees⁵⁶. The main idea is that each subsequence model intends to improve the previous model's performance by reducing errors using the gradient in a process known as boosting. The gradient minimizes the loss function by reducing the cumulative predicted errors by adding weak learners, typically DT⁵³. XGBoost also has an objective function that balances model performance and the complexity of the model. This function consists of two key components: a loss term that measures prediction accuracy and a regularization term that prevents overfitting by controlling model complexity. The objective function is expressed as follows:

$$\text{obj}(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (15)$$

where l is the loss function measuring the difference between prediction \hat{y}_i and target y_i , and Ω is the regularization term controlling model complexity.

- *KNN* is an "instance-based learning", one of the simplest non-parametric machine learning models. KNN classifier is widely used in multiple applications that include recognition and estimation and is the preferred classifier for its high simplicity and convergence²². During the inference phase, all the training values are used to assign a label to the new instance; typically, it's a memory-based learning algorithm. The algorithm assigns a label to the new instance based on the majority vote for the nearest/closest k points in the training set using similarity measures functions such as Euclidean distance, Hamming distance, and Manhattan distance, etc^{49,53,57}. In KNN, Euclidean distance is usually used for continuous variables while hamming distance is preferred for discrete variables²². For instance, the Euclidean distance is mathematically expressed as follows:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^N (x_{i,n} - x_{j,n})^2} \quad (16)$$

where x_i and x_j represent feature vectors, and N is the number of features.

- *LR* is a probabilistic-based statistical model for binomial/binary outcomes and was introduced by COX in 1958²². It uses a logistic probability distribution to model the relationship between the dependent (categorical) and independent variables. Logistic regression is easy to implement, interpretable, and provides probabilistic predictions, making it a preferred choice for binary classification. However, it assumes a linear relationship between predictors and log-odds, which can limit its effectiveness with complex or nonlinear data structures⁵⁸. LR can suffer from overfitting, especially in the high-dimensional dataset; therefore, regularization techniques such as L1-penalty and L2-penalty are used^{49,53}. The mathematical equation for LR is as follows:

$$g(z) = \frac{1}{1 + e^{-z}} \quad (17)$$

Where z is a linear regression equation

- *MLP* is a feed-forward artificial neural network model consisting of fully connected layers: an input layer (defined by the number of features), hidden layers (to learn the nonlinear representation of the input features), and an output layer (task-specific). Learning in MLP occurs by adjusting the connection weights in a back-propagation algorithm when the actual output deviates from the expected output. MLP is a commonly used supervised learning algorithm used in applications such as healthcare, finance, transportation, fitness and energy⁵⁹. MLP is capable of learning both linear and non-linear functions making it universal, it has adaptive learning properties and is able to handle complex optimization problems. Its limitations include having too many parameters, which can result in redundancy, and has a weak generalization ability for neutral problems. The main building blocks of the MLP model are neurons, weights, activation functions, loss functions, and optimizers. Neurons are the computation units that take the weighted input values and produce the nonlinear output via the activation function; weights are parametric values that the model is trained to learn (similar to learning regression coefficients). The activation function is the transfer function that introduces the nonlinearity to learn complex decision boundaries in the model; the loss function is the cost function used to quantify the errors during the forward pass. Optimizer is the mechanism to adjust the network weights during back propagation^{49,53}. Each node in the MLP incorporates a bias term, where the network processes n input variables $X = \{x_1, x_2, \dots, x_n\}$ through the input layer and produces m output variables $Y = \{y_1, y_2, \dots, y_m\}$ at the output layer⁶⁰. The MLP's total parameter count is calculated as follows:

$$n_{h1} = \sum_{k=1}^{N_{h1}} h_k h_{k+1} + h_{N_h} n \quad (18)$$

where the number of hidden nodes h_i in the i th layer is N_h . Longer computational times are required to optimize an MLP when N_h and h_k are higher.

- *RF* is an ensemble tree-based model introduced in 2001 that simultaneously trains collections of T trees (forest) independently. It uses two methods, random space approach and bagging DTs to create classification and Regression Trees (CART). These CART trees are binary decision trees built by continuously splitting data from a root node into child nodes, with each tree trained on bootstrap samples of the original data and searching randomly selected input variables for splits. This method handles challenges such as overfitting, underfitting, noise and outliers making it perfect for medical datasets²². Moreover, RF has key features that include effectively estimating missing data, managing unbalanced data accuracy by using Weighted Random Forest (WRF) and calculating variable importance in classification²². Each tree is built using a random sample or different sub-sample from the training set; the final prediction of the model denoted by \hat{y} is made based on the majority voting or average of generated tree decisions^{49,53} as provided by the following equation:

$$\hat{y} = \operatorname{argmax}_c \sum_{t=1}^T f(h_t(x) = c) \quad (19)$$

where c is the label, T is the total number of trees in the forest, $h_t(x)$ is the prediction output of the t th tree for input x .

- *SVC* is a supervised machine learning algorithm that was introduced by Vladimir Vapnik as part of his work on statistical modelling theories and methods to minimize prediction errors²². It is a powerful parametric machine learning model that uses the kernel trick to deal with nonlinear representation in the input spaces; the kernel trick transfers the data points from lower to higher multidimensional space. The latter aims to find a

good decision plane (hyperplane) that separates the different classes. SVC has been used for medical diagnosis classification and many more applications that require pattern recognition and regression estimation²². The model objective is to maximize the margin between the data points of different classes and the hyperplane—the closest points are the ones defined by the location of the hyperplane, and those points are named support vectors. The wider the margin range, the better the model. Allowing some misclassification can help to learn a wider range of margins and better hyperplane^{49,53,57}. The SVM optimization problem can be expressed as:

$$\min \frac{1}{2} \|w\|^2 \quad (20)$$

Here, y_i represents the class label (positive or negative) of a training sample i , and x_i is its feature vector representation. The optimal hyperplane is derived using the following equation:

$$y_i(w \cdot x_i + b) \geq 1, \quad \forall i \quad (21)$$

For all components of the training set, w and b must satisfy the constraints:

$$y_i(w \cdot x_i + b) = 1 \quad (22)$$

The data points x_i that satisfy $|y_i| = 1$ are identified as support vectors.

From a mathematical perspective, SVC aims to find the hyperparameter by:

$$\text{minimizing } \frac{1}{m} + C \sum p_i \quad (23)$$

where m is the margin width; p Points penalty, C is a regularization parameter (a trade-off between misclassification and margin width).

Benchmarking

We used logistic regression, which is the simplest classifier among all machine learning, as a benchmark model. We also trained the nine models using the original/all features and compared the results after applying different feature selection algorithms. In this study, we ended up training and testing 54 models (each model was trained 5 times for the outer loop, 5 times for the inner loop, and 100 times for hyperparameters optimization).

Performance evaluation metrics

We evaluated and compared the performance of different machine learning models on the testing set (unseen set) using various metrics, specifically, accuracy, Area Under the Receiver Operating Characteristics (ROC) Curve (AUC), F1-score, MCC, NPV, PPV, Specificity, and Sensitivity⁴⁷. Metrics such as NPV, PPV, specificity, and sensitivity are particularly informative when evaluating models on imbalanced datasets, such as those commonly encountered in medical applications⁶¹.

Accuracy is measuring the percentage of the predicted samples the model got right; its values range between 0% (bad model) or 0 to 100 % (perfect model) or 1, which is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (24)$$

Accuracy alone cannot be considered a good measure when working with an imbalanced dataset. Therefore, other evaluation metrics such as AUC, F1-score, and MCC must be considered. AUC is a well-known evaluation metric in the medical domain used to evaluate the discriminative capabilities of the model. AUC values range between 0 and 1; 0.5 indicates that the model made the decision based on random guessing, while 1 indicates a preferred model⁶². F1-score is a measure of accuracy; it is the harmonic mean of the precision and the recall; the measured value ranges between 0 (bad model) and 1 (perfect model). F1-measure mathematically defined as follows:

$$\text{F1 - score} = 2 * \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (25)$$

Where,

$$\text{Precision} = \frac{TP}{TP + FP} \quad (26)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (27)$$

Where, TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

MCC is a statistical measure to quantify the model's performance on all confusion matrix categories (TP, TN, FP, FN). It can be considered a Pearson correlation for discretization variables; MCC values range from -1 (bad model) to 1 (good model)^{63,64}. The MCC computed as:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (28)$$

Where, TP, TN, FP, and FN refer to true positive, true negative, false positive, and false negative, respectively.

Positive Predictive Value (PPV), also called precision, measures the proportion of positive predictions that are actually positive⁶⁵. The formula is given by:

$$PPV = \frac{TP}{TP + FP} \quad (29)$$

Where TP, TN, and FP refer to true positive, true negative, and false positive, respectively.

Negative Predictive Value (NPV) measures the proportion of negative predictions that are actually negative⁶⁵. The formula is given by:

$$NPV = \frac{TN}{TN + FN} \quad (30)$$

Where TP, TN, and FN refer to true positive, true negative, and false negative, respectively.

Sensitivity (SE) (also known as True Positive Rate or recall) quantifies the proportion of actual positives that are correctly identified⁶⁶. The formula is given by:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (31)$$

Where TP, and FN refer to true positive, and false negative, respectively.

Specificity (SP) (also known as True Negative Rate) quantifies the proportion of actual negatives that are correctly identified⁶⁶. The formula is given by:

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (32)$$

Where TN, and FP refer to true negative and false positive, respectively.

Model interpretation/explanation

Machine learning techniques provide a promising tool to tackle healthcare challenges; however, one of the main limitations of the wider use of ML in the healthcare system is the lack of model explainability and interpretability that signifies why the model made a specific decision about a particular patient⁶². For this study, we employed SHAP v0.43.0 and LIME v0.2.0.1 feature contribution analysis methods to understand and explain the output of the best predictive model. SHAP computed Shapley values which measure the contribution of each variable/feature to the final model output/prediction^{47,62}. Using Shapley values, we compute a variable importance plot for overall model analysis⁴⁵. It is important to remember that the SHAP value is interpreted as an accumulative effect of feature interaction; therefore, we can interpret it as a direct effect of a single feature⁴⁷. On the other hand, LIME is a local explanation technique that generates simplified approximations of complex models around individual predictions⁶⁷. This approach provides insight into how individual features contribute to specific predictions by perturbing the input data and observing changes in the model's output.

Data availability

The data that support the findings of this study are available from Department of Health (DOH), Abu Dhabi, UAE medical.research@doh.gov.ae but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Department of Health (DOH), Abu Dhabi.

Received: 22 July 2024; Accepted: 28 July 2025

Published online: 06 August 2025

References

1. Endocrinology, T. L. D. Spotlight on rare diseases. *Lancet Diabet. Endocrinol.* **7**, 75. [https://doi.org/10.1016/s2213-8587\(19\)30006-3](https://doi.org/10.1016/s2213-8587(19)30006-3) (2019).
2. Challa, A. P. et al. Human and machine intelligence together drive drug repurposing in rare diseases. *Front. Genet.* **12**, 836. <https://doi.org/10.3389/fgene.2021.707836> (2021).
3. Navarrete-Opazo, A. A., Singh, M., Tisdale, A., Cuttillo, C. M. & Garrison, S. R. Can you hear us now? the impact of health-care utilization by rare disease patients in the united states. *Genet. Med.* **23**, 2194–2201. <https://doi.org/10.1038/s41436-021-01241-7> (2021).
4. Brasil, S. et al. Artificial intelligence (ai) in rare diseases: Is the future brighter?. *Genes* **10**, 978. <https://doi.org/10.3390/genes10120978> (2019).
5. Ainsworth, C. Rare diseases band together toward change in research. *Nat. Med.* **26**, 1496–1499. <https://doi.org/10.1038/s41591-020-1098-7> (2020).
6. Kim, M. J. et al. The korean genetic diagnosis program for rare disease phase ii: Outcomes of a 6-year national project. *Eur. J. Hum. Genet.* **31**, 1147–1153. <https://doi.org/10.1038/s41431-023-01415-8> (2023).
7. Lagorce, D. et al. Phenotypic similarity-based approach for variant prioritization for unsolved rare disease: A preliminary methodological report. *Eur. J. Hum. Genet.* **32**, 182–189. <https://doi.org/10.1038/s41431-023-01486-7> (2023).

8. Sakate, R. & Kimura, T. Drug target gene-based analyses of drug repositionability in rare and intractable diseases. *Sci. Rep.* **11**, 428. <https://doi.org/10.1038/s41598-021-91428-4> (2021).
9. White, W. A rare disease patient/caregiver perspective on fair pricing and access to gene-based therapies. *Gene Ther.* **27**, 474–481. <https://doi.org/10.1038/s41434-019-0110-7> (2019).
10. Ehsani-Moghaddam, B., Queenan, J. A., MacKenzie, J. & Birtwhistle, R. V. Mucopolysaccharidosis type ii detection by naïve bayes classifier: An example of patient classification for a rare disease using electronic medical records from the canadian primary care sentinel surveillance network. *PLoS ONE* **13**, e0209018. <https://doi.org/10.1371/journal.pone.0209018> (2018).
11. Anderson, D., Baynam, G., Blackwell, J. M. & Lassmann, T. Personalised analytics for rare disease diagnostics. *Nat. Commun.* **10**, 13345. <https://doi.org/10.1038/s41467-019-13345-5> (2019).
12. East, K. M. et al. A state-based approach to genomics for rare disease and population screening. *Genet. Med.* **23**, 777–781. <https://doi.org/10.1038/s41436-020-01034-4> (2021).
13. Buphamalai, P., Kokotovic, T., Nagy, V. & Menche, J. Network analysis reveals rare disease signatures across multiple levels of biological organization. *Nat. Commun.* **12**, 674. <https://doi.org/10.1038/s41467-021-26674-1> (2021).
14. Faviez, C. et al. Diagnosis support systems for rare diseases: A scoping review. *Orphanet J. Rare Dis.* **15**, 1374. <https://doi.org/10.1186/s13023-020-01374-z> (2020).
15. Upadhyay, S. & Hu, H.-F. A qualitative analysis of the impact of electronic health records (ehr) on healthcare quality and safety: Clinicians' lived experiences. *Health Serv. Insights* **15**, 117863292110707. <https://doi.org/10.1177/11786329211070722> (2022).
16. Bosio, M. et al. Ediva-classification and prioritization of pathogenic variants for clinical diagnostics. *Hum. Mutat.* **40**, 865–878. <https://doi.org/10.1002/humu.23772> (2019).
17. Dehiya, V., Thomas, J. & Sael, L. Impact of structural prior knowledge in snv prediction: Towards causal variant finding in rare disease. *PLoS ONE* **13**, e0204101. <https://doi.org/10.1371/journal.pone.0204101> (2018).
18. Whicher, D., Philbin, S. & Aronson, N. An overview of the impact of rare disease characteristics on research methodology. *Orphanet J. Rare Dis.* **13**, 755. <https://doi.org/10.1186/s13023-017-0755-5> (2018).
19. Muenzer, J. Overview of the mucopolysaccharidoses. *Rheumatology* **50**, v4–v12. <https://doi.org/10.1093/rheumatology/ker394> (2011).
20. Wraith, J. & Jones, S. Mucopolysaccharidosis type i. *Pediatr. Endocrinol. Rev. PER* **12**, 102–106 (2014).
21. Schneider, K. et al. *Genereviews* (University of Washington, 1993).
22. Garavand, A. et al. Efficient model for coronary artery disease diagnosis: A comparative study of several machine learning algorithms. *J. Healthc. Eng.* **2022**, 5359540. <https://doi.org/10.1155/2022/5359540> (2022).
23. Ghaderzadeh, M. et al. Artificial intelligence in drug discovery and development against antimicrobial resistance: A narrative review. *Iran. J. Med. Microbiol.* **18**, 135–147. <https://doi.org/10.30699/ijmm.18.3.135> (2024).
24. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2**, 312. <https://doi.org/10.3389/fbinf.2022.927312> (2022).
25. Carlier, A., Vasilevich, A., Marechal, M., de Boer, J. & Geris, L. In silico clinical trials for pediatric orphan diseases. *Sci. Rep.* **8**, 2465. <https://doi.org/10.1038/s41598-018-20737-y> (2018).
26. Parenti, G., Andria, G. & Ballabio, A. Lysosomal storage diseases: From pathophysiology to therapy. *Annu. Rev. Med.* **66**, 471–486. <https://doi.org/10.1146/annurev-med-122313-085916> (2015).
27. Zhou, J., Lin, J., Leung, W. T. & Wang, L. A basic understanding of mucopolysaccharidosis: Incidence, clinical features, diagnosis, and management. *Intract. Rare Dis. Res.* **9**, 1–9. <https://doi.org/10.5582/irdr.2020.01011> (2020).
28. Kobayashi, H. Recent trends in mucopolysaccharidosis research. *J. Hum. Genet.* **64**, 127–137. <https://doi.org/10.1038/s10038-018-0534-8> (2018).
29. Khan, S. A. et al. Epidemiology of mucopolysaccharidoses. *Mol. Genet. Metab.* **121**, 227–240. <https://doi.org/10.1016/j.ymgme.2017.05.016> (2017).
30. Hirst, L., Mubeen, S., Abou-Ameira, G. & Chakrapani, A. Mucopolysaccharidosis (mps): Review of the literature and case series of five pediatric dental patients. *Clin. Case Rep.* **9**, 1704–1710. <https://doi.org/10.1002/ccr3.3885> (2021).
31. James, A., Hendriksz, C. J. & Addison, O. *The Oral Health Needs of Children, Adolescents and Young Adults Affected by a Mucopolysaccharide Disorder*, 51–58 (Springer, 2011).
32. de Almeida-Barros, R. Q. et al. Online article: Oral and systemic manifestations of mucopolysaccharidosis type vi: A report of seven cases. *Quintessence Int.* **43**, 263–263 (2012).
33. Antunes, L. A. A., Nogueira, A. P. B., Castro, G. F., Ribeiro, M. G. & Souza, I. P. R. D. Dental findings and oral health status in patients with mucopolysaccharidosis: A case series. *Acta Odontol. Scand.* **71**, 157–167. <https://doi.org/10.3109/00016357.2011.654255> (2012).
34. Turra, G. S. & Schwartz, I. V. D. Evaluation of orofacial motricity in patients with mucopolysaccharidosis: A cross-sectional study. *J. Pediatr.* **85**, 254–260 (2009).
35. Muhlebach, M. S., Wooten, W. & Muenzer, J. Respiratory manifestations in mucopolysaccharidoses. *Paediatr. Respir. Rev.* **12**, 133–138. <https://doi.org/10.1016/j.prrv.2010.10.005> (2011).
36. Berger, K. I. et al. Respiratory and sleep disorders in mucopolysaccharidosis. *J. Inherit. Metab. Dis.* **36**, 201–210 (2013).
37. Bianchi, P. M., Gaini, R. & Vitale, S. Ent and mucopolysaccharidoses. *Ital. J. Pediatr.* **44**, 57–66 (2018).
38. Patel, P. et al. Growth charts for patients with hunter syndrome. *Mol. Genet. Metab. Rep.* **1**, 5–18 (2014).
39. Parini, R., Jones, S. A., Harmatz, P. R., Giugliani, R. & Mendelsohn, N. J. The natural history of growth in patients with hunter syndrome: Data from the hunter outcome survey (hos). *Mol. Genet. Metab.* **117**, 438–446 (2016).
40. Beck, M. et al. The natural history of mps i: Global perspectives from the mps i registry. *Genet. Med.* **16**, 759–765 (2014).
41. Cohn, G. M., Morin, I., Whiteman, D. A. & Investigators, H. O. S. Development of a mnemonic screening tool for identifying subjects with hunter syndrome. *Eur. J. Pediatr.* **172**, 965–970 (2013).
42. Wainer, J. & Cawley, G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. <https://doi.org/10.48550/ARXIV.1809.09446> (2018).
43. Nicora, G., Zucca, S., Limongelli, L., Bellazzi, R. & Magni, P. A machine learning approach based on acmg/amp guidelines for genomic variant classification and prioritization. *Sci. Rep.* **12**, 6547. <https://doi.org/10.1038/s41598-022-06547-3> (2022).
44. Liang, Q. et al. Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains. *NPJ Comput. Mater.* **7**, 656. <https://doi.org/10.1038/s41524-021-00656-9> (2021).
45. Elani, H. W., Batista, A. F. M., Thomson, W. M., Kawachi, I. & Chiavegatto Filho, A. D. P. Predictors of tooth loss: A machine learning approach. *PLoS ONE* **16**, e0252873. <https://doi.org/10.1371/journal.pone.0252873> (2021).
46. Elreedy, D., Atiya, A. F. & Kamalov, F. A theoretical distribution analysis of synthetic minority oversampling technique (smote) for imbalanced learning. *Mach. Learn.* **113**, 4903–4923. <https://doi.org/10.1007/s10994-022-06296-4> (2023).
47. Hindocha, S. et al. A comparison of machine learning methods for predicting recurrence and death after curative-intent radiotherapy for non-small cell lung cancer: Development and validation of multivariable clinical prediction models. *EBioMedicine* **77**, 103911. <https://doi.org/10.1016/j.ebiom.2022.103911> (2022).
48. Theng, D. & Bhojar, K. K. Feature selection techniques for machine learning: A survey of more than two decades of research. *Knowl. Inf. Syst.* **66**, 1575–1637. <https://doi.org/10.1007/s10115-023-02010-5> (2023).
49. Noroozi, Z., Orooji, A. & Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci. Rep.* **13**, 49962. <https://doi.org/10.1038/s41598-023-49962-w> (2023).
50. Elsevier. Bat Algorithm: An overview. <https://www.sciencedirect.com/topics/computer-science/bat-algorithm>.

51. Elsevier. Genetic Algorithm: An overview. <https://www.sciencedirect.com/topics/engineering/genetic-algorithm>.
52. Wang, R. Adaboost for feature selection, classification and its relation with svm, a review. *Phys. Procedia* **25**, 800–807. <https://doi.org/10.1016/j.phpro.2012.03.160> (2012).
53. Sarkar, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.* **2**, 592. <https://doi.org/10.1007/s42979-021-00592-x> (2021).
54. Song, Y. & Lu, Y. Decision tree methods: Applications for classification and prediction. *Shanghai Arch. Psychiatry* **27**, 130–135 (2015).
55. Peretz, O., Koren, M. & Koren, O. Naive bayes classifier: An ensemble procedure for recall and precision enrichment. *Eng. Appl. Artif. Intell.* **136**, 108972. <https://doi.org/10.1016/j.engappai.2024.108972> (2024).
56. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).
57. Bzdok, D., Krzywinski, M. & Altman, N. Machine learning: Supervised methods. *Nat. Methods* **15**, 5–6. <https://doi.org/10.1038/nmeth.4551> (2018).
58. Hosmer, D. W., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (Wiley, 2013).
59. Naskath, J., Sivakamasundari, G. & Begum, A. A. S. A study on different deep learning algorithms used in deep neural nets: MLP SOM and DBN. *Wirel. Pers. Commun.* **128**, 2913–2936. <https://doi.org/10.1007/s11277-022-10079-4> (2023).
60. Chan, K. Y. et al. Deep neural networks in the cloud: Review, applications, challenges and research directions. *Neurocomputing* **545**, 126327. <https://doi.org/10.1016/j.neucom.2023.126327> (2023).
61. Thölke, P. et al. Class imbalance should not throw you off balance: Choosing the right classifiers and performance metrics for brain decoding with imbalanced data. *Neuroimage* **277**, 120253. <https://doi.org/10.1016/j.neuroimage.2023.120253> (2023).
62. Stenwig, E., Salvi, G., Rossi, P. S. & Skjærvold, N. K. Comparative analysis of explainable machine learning prediction models for hospital mortality. *BMC Med. Res. Methodol.* **22**, 1540. <https://doi.org/10.1186/s12874-022-01540-w> (2022).
63. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6413. <https://doi.org/10.1186/s12864-019-6413-7> (2020).
64. Thorsen-Meyer, H.-C. et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: A retrospective study of high-frequency data in electronic patient records. *Lancet Digital Health* **2**, e179–e191. [https://doi.org/10.1016/s2589-7500\(20\)30018-2](https://doi.org/10.1016/s2589-7500(20)30018-2) (2020).
65. Altman, D. G. & Bland, J. M. Diagnostic tests 2: Predictive values. *Br. Med. J.* **309**, 102. <https://doi.org/10.1136/bmj.309.6947.102> (1994).
66. Altman, D. G. & Bland, J. M. Diagnostic tests 1: Sensitivity and specificity. *Br. Med. J.* **308**, 1552. <https://doi.org/10.1136/bmj.308.6943.1552> (1994).
67. Ribeiro, M. T., Singh, S. & Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. <https://doi.org/10.1145/2939672.2939778> (Association for Computing Machinery, 2016).

Acknowledgements

This research is supported by ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10. Khalifa University and United Arab Emirates under Award no. KU-UAEU-2023-068. We also express our profound gratitude to Sanofi for their financial support of our research (grant fund code 21M142).

Author contributions

AA, FJ acquired funding, conceptualized the study, and supervised the project. AA built the models, analyzed the data, generated visualizations, and wrote the initial manuscript. HA, AT, MH, FJ provided the interpretation and wrote the discussion. FJ provided expert opinions and verified the methods' outcome. HA, RF, NT, AT, MH, FJ edited and reviewed the manuscript. All authors reviewed the manuscript.

Funding

This research is supported by ASPIRE, the technology program management pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), via the ASPIRE Precision Medicine Research Institute Abu Dhabi (ASPIREPMRIAD) award grant number VRI-20-10. Khalifa University and United Arab Emirates under Award no. KU-UAEU-2023-068. We also express our profound gratitude to Sanofi for their financial support of our research (grant fund code 21M142).

Declarations

Ethical approval

This study was approved by the Institutional Review Board of the Department of Health, Abu Dhabi; Approval number: IRB DOH/CMDC/2021/406. The review board waived the requirement for individual informed consent. All investigators had access to only anonymized patient information. This study was performed per the relevant laws and regulations governing research in the emirates of Abu Dhabi, UAE.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A.A.

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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025