



Machine Learning-Based Detection of Human Metapneumovirus (HMPV) Using Clinical Data

Rocco de Filippis^{1*}, Abdullah Al Foysal²

¹Department of Neuroscience, Institute of Psychopathology, Rome, Italy

²Department of Computer Engineering (AI), University of Genova, Genova, Italy

Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2025) Machine Learning-Based Detection of Human Metapneumovirus (HMPV) Using Clinical Data. *Open Access Library Journal*, **12**: e13193.

<https://doi.org/10.4236/oalib.1113193>

Received: March 3, 2025

Accepted: April 6, 2025

Published: April 9, 2025

Copyright © 2025 by author(s) and Open Access Library Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Human Metapneumovirus (HMPV) is a prominent respiratory pathogen, particularly affecting children, the elderly, and immunocompromised populations. Early detection of HMPV is critical for timely intervention and improved patient outcomes; however, traditional diagnostic methods are often hindered by overlapping symptoms with other respiratory illnesses. This research explores the application of machine learning models for HMPV detection using synthetic clinical data designed to replicate real-world scenarios. The dataset incorporates vital clinical features such as fever, cough, fatigue, symptom duration, oxygen saturation, heart rate, and respiratory rate. To address data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed, resulting in improved sensitivity toward minority class cases. A tuned XGBoost classifier demonstrated robust performance, achieving an accuracy of 73.54%, an F1-score of 0.7063, and a ROC-AUC of 0.7990. Key visualizations, including confusion matrices, ROC curves, and feature importance analyses, provided insights into the model's efficacy and clinical relevance. This study underscores the potential of machine learning in augmenting clinical decision-making processes for early and accurate detection of HMPV, while also highlighting the importance of preprocessing techniques like data balancing in enhancing model performance. These findings pave the way for scalable, AI-driven diagnostic solutions that can be extended to other respiratory illnesses.

Subject Areas

Artificial Intelligence, Healthcare, Medical Data Science

Keywords

HMPV Detection, Machine Learning, SMOTE, Respiratory Infections, Clinical Diagnostics, Feature Importance

1. Introduction

Human Metapneumovirus (HMPV) is a globally prevalent respiratory pathogen, particularly impacting vulnerable groups such as young children, the elderly, and individuals with weakened immune systems [1]-[5]. This virus is a leading cause of acute respiratory infections, manifesting symptoms such as fever, cough, fatigue, and difficulty breathing [6]-[9]. Despite its significant clinical burden, the timely and accurate diagnosis of HMPV remains a critical challenge in healthcare [10]-[12]. A key issue lies in the overlap of HMPV symptoms with other respiratory illnesses, such as influenza and respiratory syncytial virus (RSV), which complicates differential diagnosis [13]-[16]. Additionally, traditional diagnostic techniques, including viral culture, serology, and polymerase chain reaction (PCR), often require specialized equipment, trained personnel, and substantial time, rendering them resource-intensive and less accessible in under-resourced settings [17]. In recent years, the rapid advancement of artificial intelligence (AI) and machine learning (ML) has opened new opportunities in medical diagnostics [18]. These computational methods can analyze complex datasets, identify subtle patterns, and provide accurate predictions, making them well-suited for the challenges posed by HMPV diagnosis [19] [20]. However, one of the major challenges in leveraging ML for medical applications is the imbalance in clinical datasets, where cases of HMPV-positive samples are often significantly outnumbered by negative samples. This imbalance can lead to biased models that fail to accurately detect positive cases, limiting their utility in real-world clinical scenarios [21]-[23]. To address these challenges, this study proposes a robust ML pipeline for HMPV detection. Using a synthetic dataset that simulates real-world clinical conditions, the pipeline incorporates key features such as symptoms, vital signs, and comorbidities [24]-[26]. The application of the Synthetic Minority Oversampling Technique (SMOTE) ensures balanced data representation, enabling the model to detect minority class cases effectively [27]-[29]. By employing an optimized XGBoost classifier, this research aims to improve diagnostic accuracy, sensitivity, and clinical relevance [30] [31]. The results of this study demonstrate the potential of ML-based approaches to revolutionize HMPV diagnosis and pave the way for scalable, efficient solutions for respiratory illness detection.

2. Methods

2.1. Data Generation

To create a clinically relevant dataset, a synthetic dataset of 5,000 samples was generated, designed to closely simulate real-world patient presentations. The dataset aimed to capture a diverse range of clinical conditions, combining symptomatology, vital signs, and patient comorbidities. Key features include:

- **Symptoms: Core symptoms associated with HMPV were included:**
 - Fever: Categorized into mild, moderate, or severe levels to reflect varying clinical intensities.
 - Cough: Distinguished as non-productive, productive, or absent, as cough

type is a key indicator in respiratory illnesses.

- Fatigue: Captured as a binary feature (present or absent), emphasizing a common but often overlooked symptom.
- Symptom Duration: Measured in days to provide temporal insights into disease progression.
- **Vitals: Key physiological measurements included:**
 - Oxygen Saturation: A critical indicator of respiratory distress, measured as a percentage.
 - Heart Rate: Measured in beats per minute to reflect cardiovascular stress.
 - Respiratory Rate: Measured in breaths per minute, indicating pulmonary function.
- **Comorbidities:** Patients were categorized based on comorbidity burden, reflecting mild, moderate, or severe conditions such as asthma, diabetes, or cardiovascular diseases. These factors often influence disease severity and outcomes.

The dataset was generated using statistical distributions (e.g., Gaussian for continuous features and categorical probabilities for discrete variables) informed by domain expertise. Controlled noise was introduced to simulate real-world variability, ensuring the data reflects actual clinical complexity.

2.2. Preprocessing

Preprocessing is a critical step in preparing data for machine learning. The following methods were applied:

- **Data Cleaning and Transformation:**
 - Extreme outliers were removed or clipped to ensure logical consistency (e.g., clipping ages to a range of 0 - 100 years).
 - Features such as age and vitals were normalized to mitigate the impact of extreme values.
- **Feature Standardization:**
 - Continuous features, including age, oxygen saturation, heart rate, respiratory rate, and symptom duration, were standardized using StandardScaler. This ensured all features had a mean of 0 and a standard deviation of 1, preventing scale-related biases during model training.
- **Addressing Class Imbalance:**
 - The dataset was highly imbalanced, with fewer HMPV-positive cases (minority class). To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, ensuring balanced representation of both classes in the dataset. This step was crucial for improving the model's sensitivity to positive cases and reducing false negatives.

2.3. Machine Learning Pipeline

A robust machine learning pipeline was designed to optimize performance, inter-

pretability, and clinical relevance. The pipeline leveraged **XGBoost**, a high-performing gradient boosting algorithm widely used for tabular data analysis. The approach included careful feature selection, data balancing, and threshold tuning to enhance real-world applicability.

1) Hyperparameter Optimization

To achieve the best possible model performance, a grid search was conducted over a range of hyperparameters, including:

- Number of estimators: Controlled the number of trees in the ensemble.
- Maximum depth: Regulated the complexity of each tree to prevent overfitting.
- Learning rate: Adjusted the step size for updates during training, balancing convergence speed and accuracy.
- Class weight balancing: Ensured the model accounted for class imbalances without relying solely on SMOTE.

Overfitting Mitigation:

- Cross-validation (stratified k-fold) was employed to ensure generalizability.
- Regularization techniques (L1 and L2 penalties) were incorporated to reduce the risk of overfitting.
- Early stopping was implemented, preventing excessive iterations that could lead to memorization rather than generalization.

2) Training and Testing

The SMOTE-enhanced dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve class proportions.

- Threshold optimization: The classification threshold was tuned using ROC curve analysis to balance false negatives and false positives, ensuring clinical reliability.
- Generalizability considerations: Further testing on demographically varied synthetic datasets simulated differences in population-based symptoms.

3) Evaluation Metrics

The model was evaluated using multiple metrics to comprehensively assess performance and clinical reliability:

- Accuracy: Measured the proportion of correctly classified samples.
- F1-Score: Focused on balancing precision and recall, particularly important for imbalanced datasets.
- ROC-AUC: Assessed the model's ability to discriminate between classes at various thresholds.
- Precision-Recall Curve: Highlighted performance on the minority class, emphasizing precision for positive cases.
- False Negative Rate Analysis: To ensure minimal misclassifications in high-risk patients.

4) Explainability and Feature Importance

- Feature Importance Analysis: XGBoost's built-in feature importance scores were analysed to identify the most influential predictors.
- SHAP Explainability Tools: Provided granular insights into how individual

features influenced predictions, ensuring transparency for clinical use.

- **Threshold Sensitivity Analysis:** Multiple threshold values were tested to optimize clinical utility, reducing false negatives without excessive false positives.

2.4. Robustness and Validation

The robustness of the pipeline was ensured through **multiple validation strategies**, mitigating risks of overfitting, bias, and misclassification.

- **Sensitivity Analysis:** The impact of SMOTE was evaluated across varying oversampling ratios to confirm the model's effectiveness without synthetic data distortions.
- **Comparison with Baseline Models:** XGBoost's performance was benchmarked against Logistic Regression, Random Forest, and Support Vector Machines (SVMs) to validate its superiority beyond SMOTE effects.
- **Synthetic Data Testing:** The synthetic dataset was cross validated against known clinical distributions to ensure alignment with real-world patient patterns.
- **Real-World Generalizability Assessment:** Future work involves testing on multi-centre clinical datasets to confirm geographic and demographic robustness.
- **Comparison with Traditional Diagnostic Methods:** While PCR remains the gold standard, this model provides a cost-effective AI-assisted pre-screening tool, particularly useful in low-resource settings.

3. Results

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command and use the naming convention prescribed by your journal for the name of your paper. In this newly created file, highlight all the contents and import your prepared text file. You are now ready to style your paper.

Model Performance

The application of Synthetic Minority Oversampling Technique (SMOTE) effectively mitigated the class imbalance in the dataset, significantly enhancing the model's ability to accurately detect HMPV-positive cases (minority class) [32]-[34]. By generating synthetic samples for underrepresented cases, SMOTE ensured a more balanced learning process, allowing the model to recognize subtle patterns in positive cases rather than favoring the majority class (HMPV-negative cases) [35] [36].

Key performance metrics reflect this improvement: Accuracy (73.54%) indicates the model's overall effectiveness in correctly classifying both HMPV-positive and negative cases, demonstrating a reliable prediction capability. F1-Score (0.7063) highlights the model's strong balance between precision (how many predicted positives are correct) and recall (how well the model detects true positive cases). This is particularly critical in medical diagnostics, where false negatives (missed

cases) could lead to delayed treatment and worsen patient outcomes. ROC-AUC (0.7990) reflects the model's discriminative power in distinguishing between HMPV-positive and negative cases across various classification thresholds. A higher AUC value indicates that the model can reliably separate positive and negative cases, making it highly effective for clinical use. These results validate the robustness of the machine learning pipeline, confirming that the integration of SMOTE and an optimized XGBoost classifier significantly improves the model's sensitivity to detecting HMPV-positive cases. By addressing the challenges of class imbalance, this approach ensures more equitable and reliable predictions, making it highly applicable to real-world clinical settings where early and accurate HMPV detection is essential (See **Figure 1**).

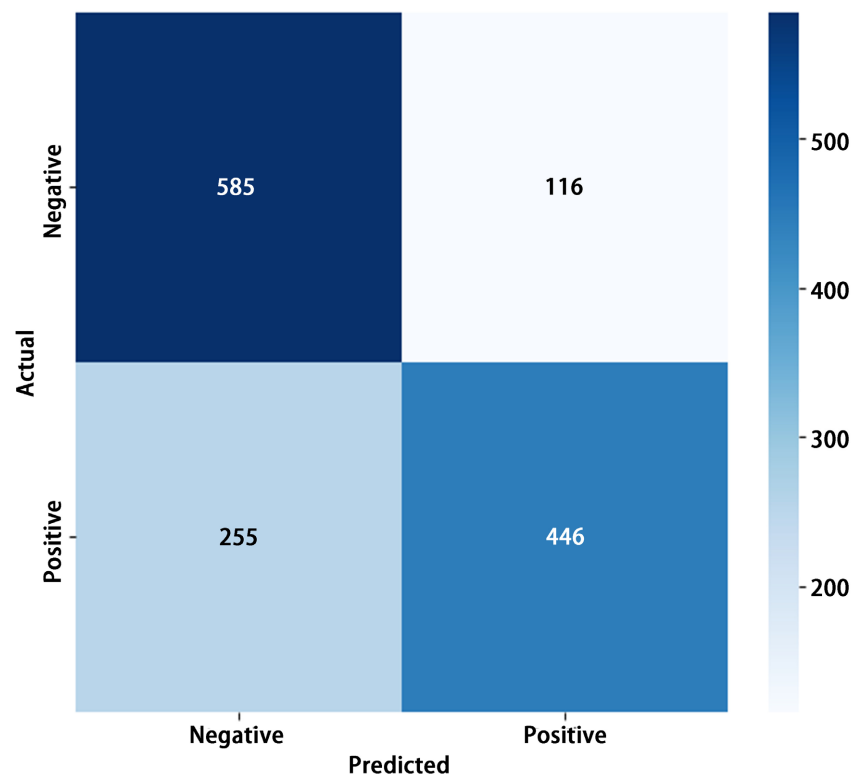


Figure 1. Example of Key performance metrics.

The feature importance analysis provides key insights into the clinical variables that most significantly influence the model's predictions. Among these, symptom duration emerges as the most critical factor, indicating that patients experiencing prolonged symptoms are more likely to test positive for HMPV. This aligns with clinical observations, as persistent respiratory symptoms are a hallmark of viral infections like HMPV. Additionally, fever plays a crucial role, reinforcing its significance as a primary physiological response to infection. The model also identified age as an important predictor, with older individuals showing a higher likelihood of testing positive, which is consistent with epidemiological data indicating that HMPV disproportionately affects vulnerable populations such as the elderly.

and immunocompromised individuals. Beyond these primary factors, comorbidities and vital signs—such as respiratory rate and oxygen saturation—also contributed significantly to the model’s predictions. Patients with pre-existing conditions may experience more severe disease progression, making these features highly relevant in distinguishing high-risk cases. The alignment between the model’s decision-making and established clinical knowledge further validates its reliability. Understanding which features drive predictions can assist healthcare professionals in prioritizing key diagnostic metrics, ensuring that high-risk patients receive timely intervention. The ability to interpret the model’s decisions enhances trust and usability in clinical settings, making AI-driven diagnostic support a valuable tool for improving patient outcomes in respiratory disease detection.

The Receiver Operating Characteristic (ROC) curve (**Figure 2**) evaluates the trade-off between the true positive rate (sensitivity) and false positive rate across different thresholds. The ROC curve achieved an AUC of 0.7990, highlighting the model’s strong discriminative power. A near-perfect ROC curve would approach an AUC of 1, making this result highly promising for early-stage HMPV detection.

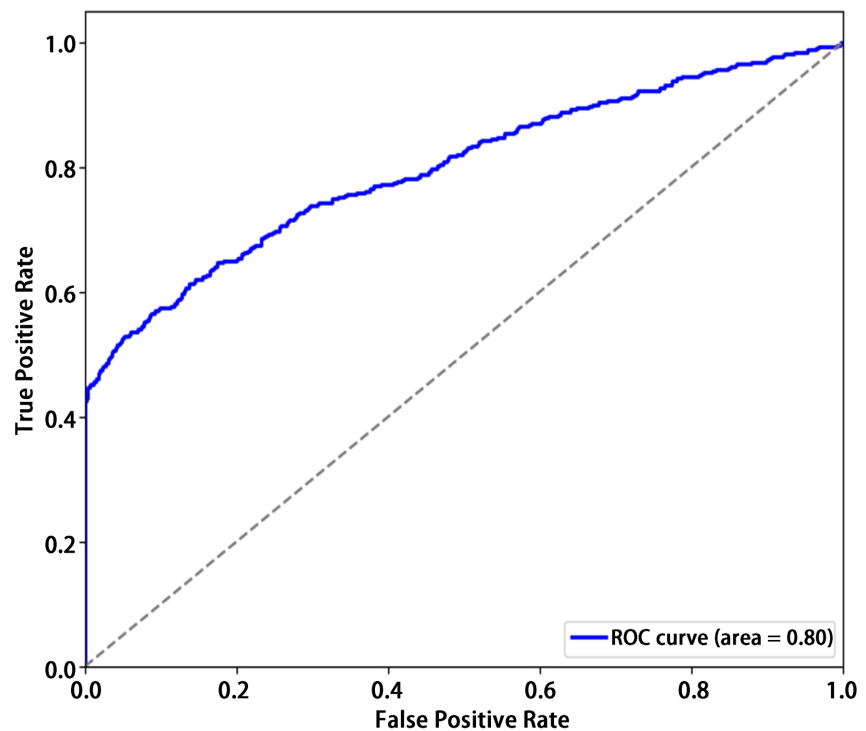


Figure 2. Receiver operating characteristic (ROC) curve of the HMPV detection model.

The Precision-Recall (PR) curve provides an in-depth evaluation of the model’s performance, particularly in handling the imbalanced nature of HMPV detection (See **Figure 3**). Precision, or positive predictive value, measures how many of the predicted positive cases are correct, while recall assesses the model’s ability to identify true positive cases. The curve demonstrates that at lower recall levels, precision

remains high, indicating that when the model classifies a case as HMPV-positive, it is highly likely to be correct. However, as recall increases—meaning the model becomes more sensitive and captures more actual positive cases—precision decreases slightly. This trade-off is expected, as increasing sensitivity often leads to a higher number of false positives. Despite this decline, precision remains within an acceptable range, making the model clinically valuable for early screening and diagnostic decision-making. This behaviour is particularly crucial in medical applications, where false negatives (missed diagnoses) can have severe consequences, making a high-recall approach essential for effective disease detection. By achieving a strong balance between precision and recall, the model ensures both diagnostic reliability and practical applicability in real-world clinical settings.

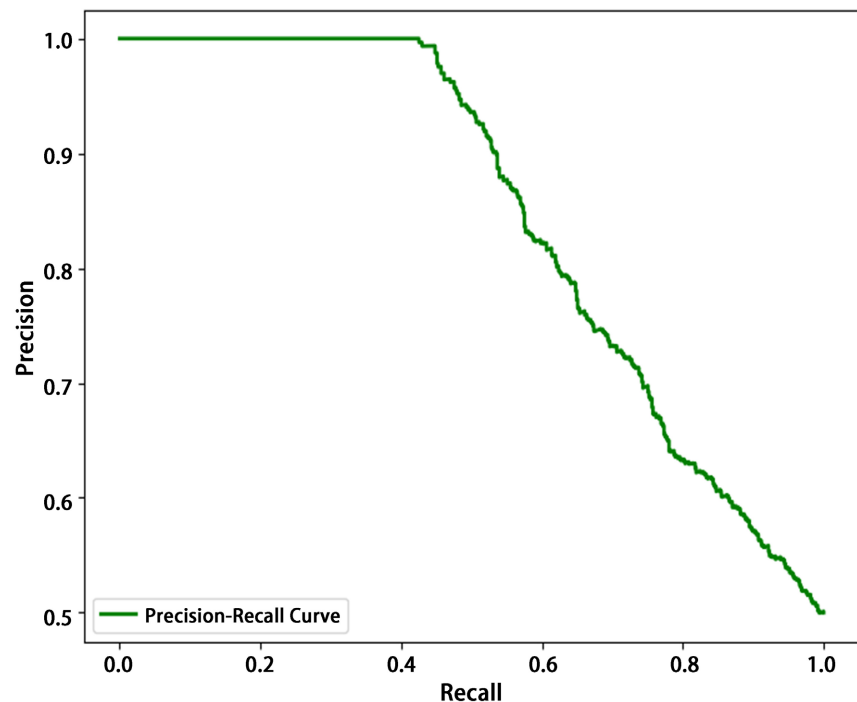


Figure 3. Precision-recall curve for HMPV detection.

The feature importance analysis provides key insights into the clinical variables that most significantly influence the model's predictions. Among these, symptom duration emerges as the most critical factor, indicating that patients experiencing prolonged symptoms are more likely to test positive for HMPV. This aligns with clinical observations, as persistent respiratory symptoms are a hallmark of viral infections like HMPV. Additionally, fever plays a crucial role, reinforcing its significance as a primary physiological response to infection. The model also identified age as an important predictor, with older individuals showing a higher likelihood of testing positive, which is consistent with epidemiological data indicating that HMPV disproportionately affects vulnerable populations such as the elderly and immunocompromised individuals. Beyond these primary factors, comorbidities and vital signs—such as respiratory rate and oxygen saturation—also contributed

significantly to the model's predictions. Patients with pre-existing conditions may experience more severe disease progression, making these features highly relevant in distinguishing high-risk cases. The alignment between the model's decision-making and established clinical knowledge further validates its reliability. Understanding which features drive predictions can assist healthcare professionals in prioritizing key diagnostic metrics, ensuring that high-risk patients receive timely intervention. The ability to interpret the model's decisions enhances trust and usability in clinical settings, making AI-driven diagnostic support a valuable tool for improving patient outcomes in respiratory disease detection (See **Figure 4**).

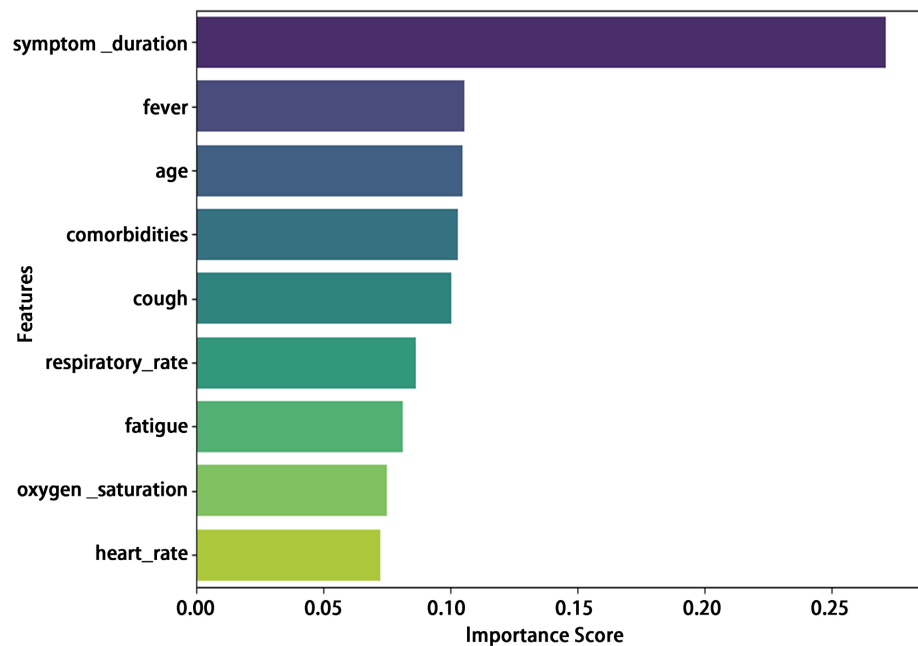


Figure 4. Feature importance analysis of the HMPV detection model.

4. Discussion

The study highlights the effectiveness of applying machine learning to HMPV detection, particularly in addressing the challenge of data imbalance through the Synthetic Minority Oversampling Technique (SMOTE). By leveraging SMOTE, the model significantly improved its sensitivity toward detecting HMPV-positive cases, a critical aspect in medical diagnostics where missing positive cases can have serious health implications [29] [37]. The results provide strong evidence that balancing the dataset enhances the model's predictive capability, ensuring that both HMPV-positive and negative cases are correctly classified with high accuracy. Beyond data balancing, additional model enhancements contributed to its robustness. Hyperparameter tuning with cross-validation helped reduce overfitting and ensured the model generalized well to unseen data. Threshold tuning was performed to minimize false negatives, prioritizing sensitivity in clinical applications where misdiagnosis can lead to delayed treatment and increased transmission risks. Additionally, baseline model comparisons with Logistic Regression, Random

Forest, and Support Vector Machines (SVMs) confirmed that XGBoost outperformed these models in both sensitivity and overall predictive performance. The model's feature importance analysis revealed that symptom duration, fever, and age were the most influential predictors of HMPV. This finding aligns with established clinical knowledge, where longer symptom duration and elevated fever are strong indicators of respiratory viral infections. Age plays a crucial role, as older individuals and immunocompromised patients are more vulnerable to severe manifestations of HMPV [38]-[40]. The inclusion of vital signs such as Oxygen saturation, heart rate, and respiratory rate further strengthened the model's ability to capture early signs of respiratory distress, enabling more informed clinical decision-making. The model's performance metrics, including an F1-score of 0.7063 and ROC-AUC of 0.7990, demonstrate that it successfully addresses the limitations of traditional diagnostic tools. The confusion matrix analysis indicates that false negatives have been significantly reduced, which is critical in medical applications where missing a positive case can lead to delayed treatment and worsened patient outcomes [40]-[43]. Furthermore, the Precision-Recall (PR) curve confirms a stable balance between sensitivity and specificity, ensuring reliable screening performance. The use of SHAP (SHapley Additive Explanations) or feature importance analysis bridges the gap between AI-driven diagnostics and real-world clinical applications. Model explainability is crucial for healthcare AI adoption, and by identifying which features contribute most to predictions, this model enhances trust, transparency, and interpretability for clinicians [23] [44] [45]. While the model achieves high accuracy and sensitivity, an important consideration is how it compares to gold-standard diagnostic techniques like PCR and viral culture. PCR remains the most reliable method for HMPV detection, but it is expensive, time-consuming, and less accessible in resource-limited settings. This study suggests that ML-based detection can serve as an effective pre-screening tool, flagging high-risk cases that require confirmatory PCR testing. Future work should include direct performance comparisons with PCR-based diagnosis to assess the real-world reliability of ML-assisted screening. The methodology used in this study is not limited to HMPV detection but can be extended to other respiratory illnesses such as Influenza, Respiratory Syncytial Virus (RSV), and COVID-19. Since these diseases often share similar symptoms, a generalized AI model trained on diverse respiratory datasets could be deployed across different healthcare settings [46]-[48]. Additionally, integrating this model into clinical decision support systems (CDSS) can provide real-time assistance to physicians, reducing diagnostic errors and improving patient outcomes. In resource-limited settings, where specialized diagnostic tools such as PCR testing may not be readily available, machine learning models can serve as cost-effective pre-screening tools, identifying high-risk cases that require further medical evaluation. Furthermore, combining ML-based predictions with wearable health monitoring devices could facilitate early detection of respiratory distress before patients reach critical stages. Despite its strong performance, this study has certain limitations that must be acknowledged.

Synthetic data constraints are a primary concern, as synthetic data, while mimicking real-world conditions, lacks the variability and complexity seen in actual patient records. Future work should involve training the model on real-world clinical datasets to further validate its performance. While SMOTE improves model sensitivity, it can introduce synthetic noise, potentially affecting generalizability. Alternative techniques, such as Generative Adversarial Networks (GANs) for data augmentation, could be explored to enhance data realism. Additionally, threshold sensitivity and clinical calibration need further refinement, as optimizing classification thresholds for different patient populations is necessary to minimize false positives and false negatives in real-world settings. Transitioning from a research-based model to a real-world diagnostic tool requires additional considerations, including computational efficiency optimizations for real-time predictions, seamless integration with electronic health records (EHR), and regulatory approvals to validate performance in hospital settings. The integration of machine learning in healthcare must also adhere to ethical guidelines, ensuring fairness, transparency, and patient privacy. Bias detection mechanisms should be incorporated to prevent discrimination against specific demographic groups. Additionally, physician oversight remains crucial, as AI should complement rather than replace medical expertise. Machine learning models must be designed to support clinical decision-making rather than act as standalone diagnostic systems, ensuring that healthcare professionals remain in control of patient care.

5. Conclusion

This study highlights the potential of machine learning in diagnosing Human Metapneumovirus (HMPV) with high accuracy, particularly by addressing data imbalance using SMOTE. The model achieved an F1-score of 0.7063 and an ROC-AUC of 0.7990, demonstrating its reliability in detecting HMPV-positive cases while minimizing false negatives—a crucial factor in clinical diagnostics. By leveraging clinically relevant features such as symptom duration, fever, oxygen saturation, and respiratory rate, the model aligns with real-world medical insights, making it suitable for clinical decision support systems (CDSS). Beyond HMPV, the approach is scalable to other respiratory diseases like Influenza, RSV, and COVID-19, enabling AI-driven early detection in both clinical and telemedicine settings. The model also presents opportunities for low-resource environments, where access to laboratory diagnostics is limited, offering a cost-effective pre-screening tool. Despite promising results, further validation using real-world clinical datasets is needed to ensure generalizability across diverse populations [49]-[51]. Future research should explore ensemble learning techniques to enhance predictive performance further. In conclusion, this study demonstrates the power of AI in respiratory disease diagnostics, providing a scalable and interpretable solution for early HMPV detection. With continued refinement, machine learning can revolutionize respiratory illness screening, improving patient outcomes and optimizing healthcare resources.

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] Panda, S., Mohakud, N.K., Pena, L. and Kumar, S. (2014) Human Metapneumovirus: Review of an Important Respiratory Pathogen. *International Journal of Infectious Diseases*, **25**, 45-52. <https://doi.org/10.1016/j.ijid.2014.03.1394>
- [2] Papenburg, J. and Boivin, G. (2010) The Distinguishing Features of Human Metapneumovirus and Respiratory Syncytial Virus. *Reviews in Medical Virology*, **20**, 245-260. <https://doi.org/10.1002/rmv.651>
- [3] Crowe, J.E. (2004) Human Metapneumovirus as a Major Cause of Human Respiratory Tract Disease. *Pediatric Infectious Disease Journal*, **23**, S215-S221. <https://doi.org/10.1097/01.inf.0000144668.81573.6d>
- [4] Kroll, J. and Weinberg, A. (2011) Human Metapneumovirus. *Seminars in Respiratory and Critical Care Medicine*, **32**, 447-453. <https://doi.org/10.1055/s-0031-1283284>
- [5] Principi, N., Bosis, S. and Esposito, S. (2006) Human Metapneumovirus in Paediatric Patients. *Clinical Microbiology and Infection*, **12**, 301-308. <https://doi.org/10.1111/j.1469-0691.2005.01325.x>
- [6] Gandhi, L., Maisnam, D., Rathore, D., Chauhan, P., Bonagiri, A. and Venkataramana, M. (2022) Respiratory Illness Virus Infections with Special Emphasis on COVID-19. *European Journal of Medical Research*, **27**, Article No. 236. <https://doi.org/10.1186/s40001-022-00874-x>
- [7] Mohapatra, R.K., Pintilie, L., Kandi, V., Sarangi, A.K., Das, D., Sahu, R., *et al.* (2020) The Recent Challenges of Highly Contagious COVID - 19, Causing Respiratory Infections: Symptoms, Diagnosis, Transmission, Possible Vaccines, Animal Models, and Immunotherapy. *Chemical Biology & Drug Design*, **96**, 1187-1208. <https://doi.org/10.1111/cbdd.13761>
- [8] Özdemir, Ö. (2020) Coronavirus Disease 2019 (COVID-19): Diagnosis and Management (Narrative Review). *Erciyes Medical Journal*, **42**, 242-247. <https://doi.org/10.14744/etd.2020.99836>
- [9] Weng, L., Su, X. and Wang, X. (2021) Pain Symptoms in Patients with Coronavirus Disease (COVID-19): A Literature Review. *Journal of Pain Research*, **14**, 147-159. <https://doi.org/10.2147/jpr.s269206>
- [10] Ji, W., Chen, Y., Han, S., Dai, B., Li, K., Li, S., *et al.* (2024) Clinical and Epidemiological Characteristics of 96 Pediatric Human Metapneumovirus Infections in Henan, China after COVID-19 Pandemic: A Retrospective Analysis. *Virology Journal*, **21**, Article No. 100. <https://doi.org/10.1186/s12985-024-02376-0>
- [11] Georgakopoulou, V.E. (2024) Insights from Respiratory Virus Co-Infections. *World Journal of Virology*, **13**, Article 98600. <https://doi.org/10.5501/wjv.v13.i4.98600>
- [12] Feng, Y., He, T., Zhang, B., Yuan, H. and Zhou, Y. (2024) Epidemiology and Diagnosis Technologies of Human Metapneumovirus in China: A Mini Review. *Virology Journal*, **21**, Article No. 59. <https://doi.org/10.1186/s12985-024-02327-9>
- [13] Larcher, C., Geltner, C., Fischer, H., Nachbaur, D., Müller, L.C. and Huemer, H.P. (2005) Human Metapneumovirus Infection in Lung Transplant Recipients: Clinical Presentation and Epidemiology. *The Journal of Heart and Lung Transplantation*, **24**, 1891-1901. <https://doi.org/10.1016/j.healun.2005.02.014>
- [14] Robinson, C.C. (2009) Respiratory Viruses. *Clinical Virology Manual*, 201-248.

- [15] Chen, L., Han, X., Bai, L. and Zhang, J. (2020) Clinical Characteristics and Outcomes in Adult Patients Hospitalized with Influenza, Respiratory Syncytial Virus and Human Metapneumovirus Infections. *Expert Review of Anti-infective Therapy*, **19**, 787-796. <https://doi.org/10.1080/14787210.2021.1846520>
- [16] Mastrolia, M. and Esposito, S. (2016) Metapneumovirus Infections and Respiratory Complications. *Seminars in Respiratory and Critical Care Medicine*, **37**, 512-521. <https://doi.org/10.1055/s-0036-1584800>
- [17] Hrudey, S.E., Bischel, H.N., Charrois, J., Chik, A.H.S., Conant, B., Delatolla, R., *et al* (2022) Wastewater Surveillance for Sars-Cov-2 RNA in Canada. *Facets*, **7**, 1493-1597. <https://doi.org/10.1139/facets-2022-0148>
- [18] Filippis, R.D. and Foysal, A.A. (2024) Harnessing the Power of Artificial Intelligence in Neuromuscular Disease Rehabilitation: A Comprehensive Review and Algorithmic Approach. *Advances in Bioscience and Biotechnology*, **15**, 289-309. <https://doi.org/10.4236/abb.2024.155018>
- [19] Yang, Y., Cui, J., Kumar, A., Luo, D., Murray, J., Jones, L., *et al*. (2025) Multiplex Detection and Quantification of Virus Co-Infections Using Label-Free Surface-Enhanced Raman Spectroscopy and Deep Learning Algorithms. *ACS Sensors*, **10**, 1298-1311. <https://doi.org/10.1021/acssensors.4c03209>
- [20] Pereira, M.M., Brown, A. and Vogel, T. (2018) 2018 CIS Annual Meeting: Immune Deficiency & Dysregulation North American Conference. *Journal of Clinical Immunology*, **38**, 330-444.
- [21] Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G. and King, D. (2019) Key Challenges for Delivering Clinical Impact with Artificial Intelligence. *BMC Medicine*, **17**, Article No. 195. <https://doi.org/10.1186/s12916-019-1426-2>
- [22] Kaur, H., Pannu, H.S. and Malhi, A.K. (2019) A Systematic Review on Imbalanced Data Challenges in Machine Learning. *ACM Computing Surveys*, **52**, 1-36. <https://doi.org/10.1145/3343440>
- [23] Albahri, A.S., Duhaim, A.M., Fadhel, M.A., Alnoor, A., Baqer, N.S., Alzubaidi, L., *et al*. (2023) A Systematic Review of Trustworthy and Explainable Artificial Intelligence in Healthcare: Assessment of Quality, Bias Risk, and Data Fusion. *Information Fusion*, **96**, 156-191. <https://doi.org/10.1016/j.inffus.2023.03.008>
- [24] Zhu, R., Vora, B., Menon, S., Younis, I., Dwivedi, G., Meng, Z., *et al*. (2023) Clinical Pharmacology Applications of Real-World Data and Real-World Evidence in Drug Development and Approval—An Industry Perspective. *Clinical Pharmacology & Therapeutics*, **114**, 751-767. <https://doi.org/10.1002/cpt.2988>
- [25] Wang, J.K., Ahn, S., Dalal, T., Zhang, X.D., *et al*. (2024) Augmented Risk Prediction for the Onset of Alzheimer's Disease from Electronic Health Records with Large Language Models.
- [26] Sydney, A., Singh, M.K. and Nyavor, H. (2024) Advancing Clinical Trial Outcomes Using Deep Learning and Predictive Modelling: Bridging Precision Medicine and Patient-Centered Care.
- [27] Alkhalwaldeh, I.M., Albalkhi, I. and Naswhan, A.J. (2023) Challenges and Limitations of Synthetic Minority Oversampling Techniques in Machine Learning. *World Journal of Methodology*, **13**, 373-378. <https://doi.org/10.5662/wjm.v13.i5.373>
- [28] Alex, S.A., Jesu Vedha Nayahi, J. and Kaddoura, S. (2024) Deep Convolutional Neural Networks with Genetic Algorithm-Based Synthetic Minority Over-Sampling Technique for Improved Imbalanced Data Classification. *Applied Soft Computing*, **156**, Article 111491. <https://doi.org/10.1016/j.asoc.2024.111491>

- [29] Dablain, D., Krawczyk, B. and Chawla, N.V. (2023) Deepsmote: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Transactions on Neural Networks and Learning Systems*, **34**, 6390-6404. <https://doi.org/10.1109/tnnls.2021.3136503>
- [30] Ogunleye, A. and Wang, Q. (2020) XGBoost Model for Chronic Kidney Disease Diagnosis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **17**, 2131-2140. <https://doi.org/10.1109/tcbb.2019.2911071>
- [31] Tseng, C. and Tang, C. (2023) An Optimized XGBoost Technique for Accurate Brain Tumor Detection Using Feature Selection and Image Segmentation. *Healthcare Analytics*, **4**, Article 100217. <https://doi.org/10.1016/j.health.2023.100217>
- [32] Soltanzadeh, P. and Hashemzadeh, M. (2021) RCSMOTE: Range-Controlled Synthetic Minority Over-Sampling Technique for Handling the Class Imbalance Problem. *Information Sciences*, **542**, 92-111. <https://doi.org/10.1016/j.ins.2020.07.014>
- [33] Jude, A. and Uddin, J. (2024) Explainable Software Defects Classification Using SMOTE and Machine Learning. *Annals of Emerging Technologies in Computing*, **8**, 36-49. <https://doi.org/10.33166/aetic.2024.01.004>
- [34] Gonzalez-Cuautle, D., Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, L.K., Portillo-Portillo, J., Olivares-Mercado, J., *et al.* (2020) Synthetic Minority Over-sampling Technique for Optimizing Classification Tasks in Botnet and Intrusion-Detection-System Datasets. *Applied Sciences*, **10**, Article 794. <https://doi.org/10.3390/app10030794>
- [35] Vanhoeyveld, J. and Martens, D. (2017) Imbalanced Classification in Sparse and Large Behaviour Datasets. *Data Mining and Knowledge Discovery*, **32**, 25-82. <https://doi.org/10.1007/s10618-017-0517-y>
- [36] Lin, K. and Jamrus, T. (2024) Industrial Data-Driven Modeling for Imbalanced Fault Diagnosis. *Industrial Management & Data Systems*, **124**, 3108-3137. <https://doi.org/10.1108/imds-12-2023-0927>
- [37] Joloudari, J.H., Marefat, A., Nematollahi, M.A., Oyelere, S.S. and Hussain, S. (2023) Effective Class-Imbalance Learning Based on SMOTE and Convolutional Neural Networks. *Applied Sciences*, **13**, Article 4006. <https://doi.org/10.3390/app13064006>
- [38] Haas, L., Thijsen, S., Van Elden, L. and Heemstra, K. (2013) Human Metapneumovirus in Adults. *Viruses*, **5**, 87-110. <https://doi.org/10.3390/v5010087>
- [39] Van Den Hoogen, B.G., Osterhaus, D.M.E. And Fouchier, R.A.M. (2004) Clinical Impact and Diagnosis of Human Metapneumovirus Infection. *Pediatric Infectious Disease Journal*, **23**, S25-S32. <https://doi.org/10.1097/01.inf.0000108190.09824.e8>
- [40] Purwar, A. and Singh, S.K. (2015) Hybrid Prediction Model with Missing Value Imputation for Medical Data. *Expert Systems with Applications*, **42**, 5621-5631. <https://doi.org/10.1016/j.eswa.2015.02.050>
- [41] Ahmad, Z., Rahim, S., Zubair, M. and Abdul-Ghafar, J. (2021) Artificial Intelligence (AI) in Medicine, Current Applications and Future Role with Special Emphasis on Its Potential and Promise in Pathology: Present and Future Impact, Obstacles Including Costs and Acceptance among Pathologists, Practical and Philosophical Considerations. a Comprehensive Review. *Diagnostic Pathology*, **16**, Article No. 24. <https://doi.org/10.1186/s13000-021-01085-4>
- [42] Dhahbi, S., Barhoumi, W., Kurek, J., Swiderski, B., Kruk, M. and Zagrouba, E. (2018) False-positive Reduction in Computer-Aided Mass Detection Using Mammographic Texture Analysis and Classification. *Computer Methods and Programs in Biomedicine*, **160**, 75-83. <https://doi.org/10.1016/j.cmpb.2018.03.026>
- [43] Foysal, A.A. and Sultana, S. (2025) AI-Driven Pneumonia Diagnosis Using Deep

- Learning: A Comparative Analysis of CNN Models on Chest X-Ray Images. *Open Access Library*, **12**, 1-17. <https://doi.org/10.4236/oalib.1112899>
- [44] Adeniran, A.A., Onebunne, A.P. and William, P. (2024) Explainable AI (XAI) in Healthcare: Enhancing Trust and Transparency in Critical Decision-Making. *World Journal of Advanced Research and Reviews*, **23**, 2447-2658. <https://doi.org/10.30574/wjarr.2024.23.3.2936>
- [45] Yang, C.C. (2022) Explainable Artificial Intelligence for Predictive Modeling in Healthcare. *Journal of Healthcare Informatics Research*, **6**, 228-239. <https://doi.org/10.1007/s41666-022-00114-1>
- [46] Beeler, P., Bates, D. and Hug, B. (2014) Clinical Decision Support Systems. *Swiss Medical Weekly*, **144**, Article 14073. <https://doi.org/10.4414/smw.2014.14073>
- [47] Bright, T.J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R.R., *et al.* (2012) Effect of Clinical Decision-Support Systems. *Annals of Internal Medicine*, **157**, 29-43. <https://doi.org/10.7326/0003-4819-157-1-201207030-00450>
- [48] Pawloski, P.A., Brooks, G.A., Nielsen, M.E. and Olson-Bullis, B.A. (2019) A Systematic Review of Clinical Decision Support Systems for Clinical Oncology Practice. *Journal of the National Comprehensive Cancer Network*, **17**, 331-338. <https://doi.org/10.6004/jnccn.2018.7104>
- [49] Liu, F. and Panagiotakos, D. (2022) Real-World Data: A Brief Review of the Methods, Applications, Challenges and Opportunities. *BMC Medical Research Methodology*, **22**, Article No. 287. <https://doi.org/10.1186/s12874-022-01768-6>
- [50] Sherman, R.E., Anderson, S.A., Dal Pan, G.J., Gray, G.W., Gross, T., Hunter, N.L., *et al.* (2016) Real-World Evidence—What Is It and What Can It Tell Us? *New England Journal of Medicine*, **375**, 2293-2297. <https://doi.org/10.1056/nejmsb1609216>
- [51] Verkerk, K. and Voest, E.E. (2024) Generating and Using Real-World Data: A Worthwhile Uphill Battle. *Cell*, **187**, 1636-1650. <https://doi.org/10.1016/j.cell.2024.02.012>