Proceedings of the International Conference on Sustainable Expert Systems (ICSES-2024)
IEEE Xplore Part Number: CFP24VS6-ART; ISBN: 979-8-3315-4036-4

Machine Learning in Healthcare: Decision Trees for Asthma Risk Prediction

Tanishq Soni
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
tanishq.soni@chitkara.edu.in

Deepali Gupta
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
deepali.gupta@chitkara.edu.in

Monica Dutta
Chitkara University Institute of Engineering and Technology, Chitkara University
Punjab, India
monica.dutta@chitkara.edu.in

2024 4th International Conference on Sustainable Expert Systems (ICSES) | 979-8-3315-4036-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSES63445.2024.10763341

***Abstract*—A chronic respiratory condition marked by hyperreactivity and inflammation of the airways, asthma presents serious health problems worldwide. Asthma prediction done early and precisely can result in better patient outcomes and care. The effectiveness of many machine learning algorithms in asthma prediction is investigated in this work, with an emphasis on a performance comparison of Decision Tree, K-Nearest Neighbours (KNN), and Random Forest classifiers. Created prediction algorithms to detect people at risk of asthma using a large dataset including clinical, environmental, and genetic variables. Among KNN and Random Forest classifiers, the Decision Tree method outperforms them with the maximum prediction accuracy of 81%. The Decision Tree model outperforms others because of its interpretability, which offers precise understanding of the decision-making process, and its capacity to manage complicated relationships between elements. The possibilities of Decision Tree models in asthma prediction are demonstrated by these results, which also emphasize the need of choosing suitable machine learning methods for efficient illness prediction. This work offers a potential method for early asthma diagnosis and customized therapeutic techniques, therefore supporting the continuous attempts to use machine learning in healthcare.**

Keywords—Asthma, Decision Tree, KNN, Random Forest, Machine Learning

I. INTRODUCTION

Breathing becomes difficult with asthma, a chronic respiratory disease marked by inflammation and airway constriction [1]. Frequently brought on by different allergens, irritants, or physical exercise, common symptoms include wheezing, dyspnoea, chest tightness, and coughing. Because of the inflammation, the airways become extremely sensitive, and triggers can cause an asthma attack, in which the symptoms sharply intensify and need for quick treatment [2]. Though it can strike anybody at any age, asthma usually begins in infancy. According to the World Health Organisation, asthma is thought to impact 262 million people worldwide and result in 461,000 deaths yearly [3]. Asthma incidence varies greatly among nations and is impacted by a confluence of environmental and genetic variables. Asthma management is avoiding recognised causes and controlling symptoms with drugs like bronchodilators and inhaled corticosteroids [4].

There is a complicated interaction between a genetic predisposition and environmental exposures that culminates in the development of asthma. Individuals who have a family history of asthma or other allergy illnesses that are connected to asthma, such as eczema and hay fever, are at a greater risk of acquiring asthma, which indicates that there is a strong genetic component behind the condition [5]. Additionally, this propensity frequently interacts with environmental variables that have the potential to either initiate or intensify the illness. The allergens pollen, pet dander, dust mites, and mould are examples of common environmental triggers that can cause allergic reactions [6]. Infections in the respiratory system, particularly those that occur during the early years of life, can also play a significant impact by causing damage to the delicate tissue of the lungs and perhaps leading to persistent asthma. In addition, being exposed to air pollutants such as tobacco smoke, industrial pollution, and exhaust from vehicles can irritate the airways and make asthma symptoms even more severe [7]. There are a number of other elements that can play a role in the development and progression of asthma symptoms. These elements include occupational risks (such as being exposed to dust or chemicals in the job), harsh weather conditions, and even stress [8]. Asthma is a very prevalent and chronic health concern all over the world due to the number of factors that contribute to its development.

The term "machine learning" (ML) refers to the process of analysing big datasets in order to recognise patterns that may not be obvious using conventional approaches [9]. This has the potential to significantly improve the prognosis and treatment of asthma. To construct prediction models that are very accurate in identifying individuals who are at a high risk of getting asthma, machine learning algorithms may integrate a wide variety of data sources, such as electronic health records, genetic information, environmental data, and patient-reported symptoms [10]. The intensity of asthma episodes and the potential factors that cause them may also be predicted by these models, which enables more individualised and preventative medical care measures. In addition, machine learning may assist in the monitoring of the evolution of diseases and the success of recommended therapies, including the modification of these treatments based on real-time data inputs [11]. The ability to successfully harness big data not only helps in early diagnosis and the development of individualised treatment programmes, but it also helps in understanding the complex interactions that occur between the many risk factors connected with asthma. This, in turn, may eventually result in improved patient outcomes and lower costs associated with healthcare [12].

Authorized licensed use limited to: VIT University. Downloaded on January 05,2025 at 05:11:13 UTC from IEEE Xplore. Restrictions apply.

II. LITERATURE SURVEY

Feng et al. [1] focus on Asthma and Chronic Obstructive Pulmonary Disease," the author explores the use of artificial intelligence and machine learning techniques in the diagnosis, classification, management, and treatment of asthma and COPD disease. The socioeconomic impact and prevalence of these diseases are brought to light, particularly in nations that are still in the process of developing. Even while there are rules, the role that precision medicine plays is still very restricted. AI and ML approaches, particularly those used in genomics and medical imaging, have demonstrated that they have the potential to analyse massive amounts of medical data. The considerable clinical influence they have, however, is still very restricted. The study highlights the promise and constraints of artificial intelligence and machine learning in improving the screening, diagnosis, categorization, monitoring, and treatment of chronic airway disorders. It also advocates for next measures to guarantee that its deployment in clinical settings is both successful and safe.

Bose el al. [2] highlighted a machine learning approach used five machine learning models to forecast asthma persistence in kids diagnosed before turning five. Among these were gradient boosted trees (XGBoost), random forest, k-nearest neighbours, logistic regression, and naïve Bayes. Electronic health records for 9,934 children made up the dataset; 8802 were found to have chronic asthma and 1132 to have temporary asthma. At 0.43 as the mean average NPV-Specificity area (ANSA), the XGBoost model showed the greatest performance. With ANSA values of 0.42, other models including logistic regression and random forest also did well. Important factors found were the age at last asthma diagnosis, the total number of asthma-related visits, self-identified black race, allergic rhinitis, and eczema. These results demonstrate the ability of machine learning algorithms to identify which children will continue to have asthma symptoms, therefore supporting early childhood asthma care.

TahaSamadSoltaniHeris et al. [3] utilised several machine learning methods to asthma diagnosis. From two Tehrani hospitals, 169 asthmatics and 85 non-asthmatics made up the dataset. Random forests (RF), support vector machines (SVM) and k-nearest neighbours (KNN) were among the methods examined. With five neighbours, the KNN algorithm produced ideal results with 100% accuracy, sensitivity, and specificity. The accuracy, specificity, and sensitivity of the SVM using a radial basis function kernel were all 98.70%, 97.37%, and 99.34%. The RF approach produced 98.68% sensitivity, 92.11% specificity, and 96.52% accuracy with 20 trees. In a comparison with a fuzzy expert system and other techniques, these models showed better asthma diagnosis accuracy. The work emphasises how useful machine learning—KNN in particular—is for medical diagnosis and how crucial pre-processing and parameter optimisation are to improve model performance.

BHAT et al. [4] investigates the use of many machine learning methods in asthma diagnosis. In all, 169 asthmatics and 85 non-asthmatics from two Tehran hospitals made up the dataset. KNN, SVM, and random forests (RF) were the machine learning algorithms assessed. At 1.0 for specificity, sensitivity, and accuracy, the KNN algorithm with five neighbours outperformed the others. A radial basis function kernel SVM achieved 0.9870 accuracy, 0.9737 specificity, and 0.9934 sensitivity. The sensitivity of 0.9868, specificity of 0.9211, and accuracy of 0.9652 were obtained using the RF approach utilising 20 trees. KNN shown to perform the best overall among these models when compared to a fuzzy expert system and other diagnostic techniques. The work demonstrates how well machine learning algorithms work in medical diagnosis, highlighting in particular how crucial pre-processing and parameter optimisation are to improving model performance.

AKBAR et al. [5] analyses k-nearest neighbours (KNN), support vector machines (SVM), and random forests (RF) as machine learning methods for asthma diagnosis. From two Tehrani hospitals, 169 asthmatic and 85 non-asthmatic patients made up the sample. At values of 1.0 for accuracy, specificity, and sensitivity, the KNN algorithm with five neighbours showed the best result. A radial basis function kernel SVM produced values of 0.9870, 0.9737, and 0.9934, in that order. Results from the RF technique with 20 trees were 0.9652, 0.9211, and 0.9868. KNN shown to perform the best overall among these models when compared to a fuzzy expert system and other diagnostic techniques. The work emphasises the use of machine learning algorithms in medical diagnosis and the need of parameter optimisation and pre-processing to improve model performance.

Kothalawala et al. [6] built two machine learning models, CAPP and CAPE, to forecast childhood asthma at ten years old. Performance was better in the CAPP model (using a linear SVM with twelve variables) and the CAPE model (using a radial basis function support vector machine (RBF SVM) with eight predictors than in the conventional logistic regression models. The AUC (area under the curve) for the CAPE model was 0.71, while for the CAPP model it was 0.82. Good generalizability of both models was shown by validation using data from the Manchester Asthma and Allergy Study (MAAS). The study found important indicators including early life cough and wheeze for CAPE and preschool cough, atopy, and polysensitization for CAPP, and demonstrated the potential of machine learning to increase asthma prediction accuracy. SHAP values were used to assist explain the models' predictions, therefore removing a significant obstacle to machine learning's therapeutic use. Large datasets and outside validation are also stressed by the study as being essential to guaranteeing the accuracy and applicability of prediction models in a variety of demographics.

Exarchos et al. [7] assesses the use of artificial intelligence (AI) and machine learning (ML) to many facets of asthma study. In four primary categories—asthma screening and diagnosis, patient categorization, asthma management and monitoring, and asthma treatment—98 papers from 1988 to 2019 were included in the systematic review. Among the important AI/ML methods employed are k-nearest neighbours (k-NN), decision trees, random forests (RFs), support vector machines (SVMs), and artificial neural networks (ANNs). With the SVM, for instance, demonstrating great performance in identifying airway obstruction using forced oscillation technique (FOT) data, studies employing these techniques showed encouraging results in asthma prediction and classification. Numerous research, meanwhile, were constrained by things like tiny sample sizes and the requirement for larger data sets. Better diagnosis accuracy, patient categorization, and treatment techniques are some of the ways that AI/ML can enhance asthma care, but the study also stresses the necessity of bigger studies and more research to close present gaps and confirm the results.

Authorized licensed use limited to: VIT University. Downloaded on January 05,2025 at 05:11:13 UTC from IEEE Xplore. Restrictions apply.

Fontanella et al. [8] highlights about the developing place of artificial intelligence (AI) and machine learning (ML) in asthma research. It draws attention to the change in research methodology from conventional hypothesis-driven to data-driven approaches that reveal patterns in huge and complicated dataset. It has been difficult to use these discoveries in clinical practice even with the abundance of data and technical developments. The study groups the studies into areas including environmental exposures, atopic illnesses, respiratory disease diagnostics, and asthma heterogeneity. Improvement in diagnostic accuracy and identification of asthma subtypes and biomarkers have been demonstrated by the application of ML and AI. But because of worries about accuracy, dependability, and interpretability, these technologies are still seldom used in clinical environments. To fully use big data and contemporary analytics, the paper highlights the requirement of integrated, multidisciplinary teams that guarantee significant clinical insights and improvements in asthma care and treatment. To help better understand and treat asthma, the paper ends by arguing for a balance between explanatory and predictive models.

III. Dataset

The dataset taken from Kaggle consists of 316,800 entries. A variety of characteristics pertaining to symptoms, age, gender, and the severity of illnesses are included in the dataset collection. There are a number of features, such as Tiredness, Dry-Cough, Difficulty-in-Breathing, Sore-Throat, and None_Sympton, which indicate whether or not these particular symptoms are present. In addition, it includes characteristics such as Pains, Nasal Congestion, and Runny Nose in order to capture other symptoms that are commonly experienced. Other properties, such as None_Experiencing, are used to indicate circumstances in which the patient did not experience any symptoms. Age-related characteristics are broken down into the following age categories: Age_0-9, Age_10-19, Age_20-24, Age_25-59, and Age_60+. This allows for the classification of individuals into distinct age groups. Information pertaining to gender is included in the dataset by means of the Gender_Female and Gender_Male characteristics. Last but not least, the severity of the disease is categorised using qualities such as Severity_Mild, Severity_Moderate, and Severity_None. These qualities, when taken as a whole, offer a thorough picture of the many different elements that were taken into consideration in the dataset. The dataset contains an equal number of males and females, with each gender having 105,600 entries.

The previous work was done on the random forest on this dataset. For the 80% data is utilised for the training and remaining 20% is for testing. After the data splitting the random forest achieve the accuracy of 75%.

IV. Proposed Model

This dataset is collected from Kaggle having 316800 entries related to asthma having different attributes that can cause asthma. 2 different machine learning models are proposed i.e. KNN and decision tree. There 3 class for the asthma which are none asthma, moderate asthma and mild asthma. For the 2 different models the dataset is divided into 2. 80% dataset is used for training and 20% is for testing purpose.

Firstly the dataset was tested on KNN classifier and it achieve the accuracy of 43%. Figure 1 shows the confusion matrix for the KNN.

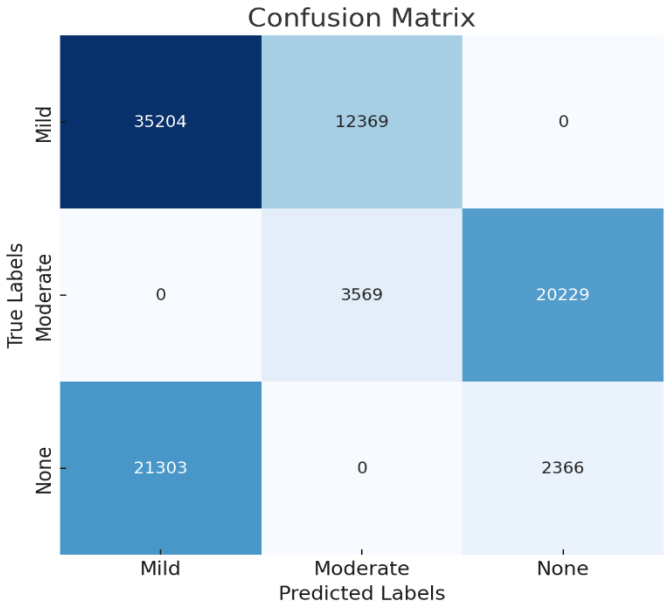

Fig. 1. Confusion Matrix of KNN Classifier

On the same dataset decision tree is applied and it achieves the highest accuracy 81% on predicting the 3 classes of the asthma. Figure 2 shows the confusion matrix for the decision tree.

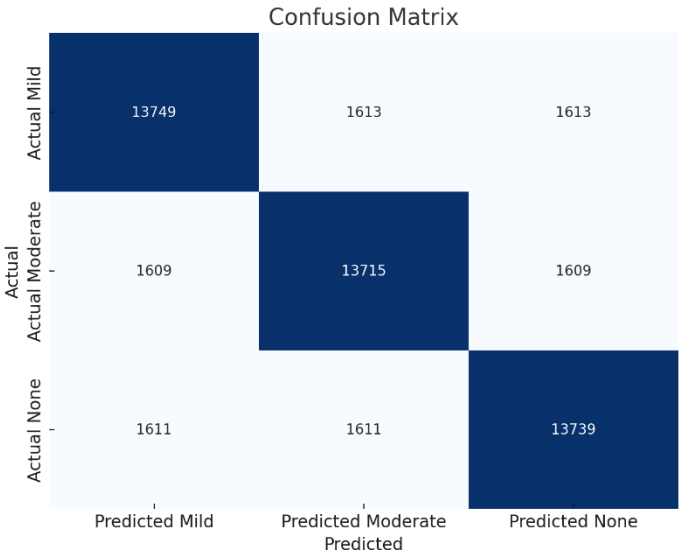

Fig. 2. Decision Tree Confusion Matrix

Figure 3 is the accuracy comparsion between the 3 different machine learning models. The models are trained and tested on the same data split. From this previous random forest have the accuracy of 75%. Tested decision tree have the highest accuracy of 81% as compared to KNN with 43%.

Authorized licensed use limited to: VIT University. Downloaded on January 05,2025 at 05:11:13 UTC from IEEE Xplore. Restrictions apply.

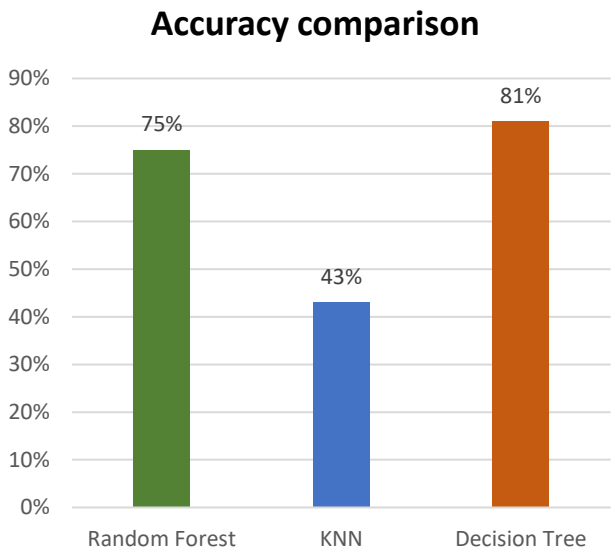

Fig. 3. Accuracy Comparison

V. CONCLUSION

Outperforming the random forest and K-Nearest Neighbours (KNN) models, the decision tree model earned the greatest accuracy of 81% based on the dataset analysis. It appears from this higher performance that the decision tree model is more appropriate for identifying the trends and connections in this dataset. Because of its form, which enables it to manage the categorical character of the data, the decision tree facilitates the interpretation and comprehension of the decision-making process. The huge and complicated structure of this dataset challenged the KNN model, which reduced accuracy even if it was helpful for basic and small datasets. Also, in this instance the performance of the single decision tree was not surpassed by the random forest model, which usually works well by aggregating several decision trees. With its high accuracy in categorising the severity of symptoms in this dataset, the decision tree model is the better option than KNN and random forest models for this particular use.

REFERENCES

[1] Feng, Y., Wang, Y., Zeng, C. and Mao, H., 2021. Artificial intelligence and machine learning in chronic airway diseases: focus on asthma and chronic obstructive pulmonary disease. International journal of medical sciences, 18(13), p.2871.

[2] Bose, S., Kenyon, C.C. and Masino, A.J., 2021. Personalized prediction of early childhood asthma persistence: a machine learning approach. PloS one, 16(3), p.e0247784.

[3] Tahasamadsoltaniheris, M., Mahmoodvand, Z. and Zolnoori, M., 2013. Intelligent diagnosis of Asthma using machine learning algorithms. International Research Journal of Applied and Basic Sciences, 5(1), pp.140-145.

[4] Bhat, G.S., Shankar, N., Kim, D., Song, D.J., Seo, S., Panahi, I.M. and Tamil, L., 2021. Machine learning-based asthma risk prediction using IoT and smartphone applications. IEEE Access, 9, pp.118708-118715.

[5] AKBAR, W., WU, W.P., FAHEEM, M., SALEEM, M.A., GOLILARZ, N.A. and HAQ, A.U., 2019, December. Machine learning classifiers for asthma disease prediction: a practical illustration. In 2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing (pp. 143-148). IEEE.

[6] Kothalawala, D.M., Murray, C.S., Simpson, A., Custovic, A., Tapper, W.J., Arshad, S.H., Holloway, J.W., Rezwan, F.I. and STELAR/UNICORN investigators, 2021. Development of childhood asthma prediction models using machine learning approaches. Clinical and Translational Allergy, 11(9), p.e12076.

[7] Exarchos, K.P., Beltsiou, M., Votti, C.A. and Kostikas, K., 2020. Artificial intelligence techniques in asthma: a systematic review and critical appraisal of the existing literature. European Respiratory Journal, 56(3).

[8] Fontanella, S., Cucco, A. and Custovic, A., 2021. Machine learning in asthma research: moving toward a more integrated approach. Expert Review of Respiratory Medicine, 15(5), pp.609-621.

[9] Soni, T., Uppal, M., Gupta, D. and Gupta, G., 2023, May. Efficient machine learning model for cardiac disease prediction. In 2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) (pp. 1-5). IEEE.

[10] Rani, S., Koundal, D., Kavita, F., Ijaz, M.F., Elhoseny, M. and Alghamdi, M.I., 2021. An optimized framework for WSN routing in the context of industry 4.0. Sensors, 21(19), p.6474.

[11] Soni, T., Gupta, D. and Uppal, M., 2023, December. Transforming the Prediction of Heart Disease: An Empirical Analysis of Machine Learning Classifiers. In 2023 IEEE Pune Section International Conference (PuneCon) (pp. 1-5). IEEE.

[12] Goyal, N., Dave, M. and Verma, A.K., 2020. SAPDA: secure authentication with protected data aggregation scheme for improving QoS in scalable and survivable UWSNs. Wireless Personal Communications, 113(1), pp.1-15.

Authorized licensed use limited to: VIT University. Downloaded on January 05,2025 at 05:11:13 UTC from IEEE Xplore. Restrictions apply.