Düzce University Journal of Science & Technology

Research Article

A New Hybrid Classification Framework in Childhoods Allergies with Dataset Slicing Method

Pınar KARADAYI ATAŞ[a,*]

[a] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İstanbul Arel Üniversitesi, Istanbul, TÜRKİYE*
* *Sorumlu yazarın e-posta adresi: pinaratas@arel.edu.tr*
DOI: 10.29130/dubited.1353771

ABSTRACT

Childhood allergies, particularly food allergies, are growing more frequent. Their major influence on children's health and well-being has piqued the interest of worldwide public health officials. The increased prevalence of childhood allergies in Turkey, where these patterns are also relevant, adds urgency to the need for effective classification and management options. This study addresses the shortcomings of simple classification algorithms in obtaining high accuracy by presenting a novel hybrid classification methodology. The research creates a novel method where three different prediction models are built by combining Support Vector Machine and Decision Tree classifiers. This method improves the classification process by taking into account instances that have been incorrectly classified as possible sources of useful information instead of just being noise. This instance filtering-based hybrid classification algorithm that is used in this study maintains the simplicity of interpreting learning outcomes while achieving comparatively high accuracy. Extensive experiments on the allergy dataset show the effectiveness of this hybrid approach, with an impressive accuracy of 0.906. This greatly outperforms the fundamental classification algorithms. The experimental outputs have important implications for medical professionals. This study might add a valuable contribution to the literature by giving a fresh solution to childhood allergy classification.

Keywords: *Childhood allergies, Hybrid classification, Machine learning*

Çocukluk Alerjilerinde Veri Kesme Yöntemiyle Yeni Bir Hibrit Sınıflandırma Çerçevesi

ÖZ

Çocukluk alerjileri, özellikle de gıda alerjileri giderek artmaktadır. Çocukların sağlığı ve refahı üzerindeki büyük etkileri dünya çapındaki halk sağlığı yetkililerinin ilgisini çekmektedir. Bu kalıpların da geçerli olduğu Türkiye'de çocukluk çağı alerjilerinin artan prevalansı, etkili sınıflandırma ve yönetim seçeneklerine olan ihtiyacın aciliyetini artırmaktadır. Bu çalışma, yeni bir hibrit sınıflandırma metodolojisi sunarak, basit sınıflandırma algoritmalarının yüksek doğruluk elde etmedeki eksikliklerini gidermektedir. Araştırma, Destek Vektör Makinesi ve Karar Ağacı sınıflandırıcılarını birleştirerek üç farklı tahmin modelinin oluşturulduğu yeni bir yöntem yaratmaktadır. Çalışmamızda kullanılan bu yöntem, yanlış sınıflandırılan örnekleri sadece gürültü olarak değil, potansiyel olarak kullanışlı bilgi kaynakları olarak ele alarak sınıflandırma sürecini geliştirir. Bu örnek filtreleme tabanlı hibrit sınıflandırma algoritması, nispeten yüksek doğruluk elde ederken öğrenme çıktılarını yorumlamanın basitliğini korur. Alerji veri seti üzerinde yapılan kapsamlı deneyler, bu hibrit yaklaşımın etkinliğini 0,906'lık etkileyici bir doğrulukla göstermektedir. Bu, temel algoritmaların yanı sıra daha önce önerilen klasik sınıflandırma algoritmalarından da büyük ölçüde daha iyi performans sergilemektedir. Deneysel çıktıların tıp uzmanları için önemli sonuçları vardır. Bu çalışma çocukluk çağı alerji sınıflandırmasına yeni bir çözüm sunarak literatüre değerli bir katkı sağlayabilir.

Received: 01/09/2023, Revised: 07/12/2023, Accepted: 11/02/2024

Anahtar Kelimeler: *Çocukluk alerjileri, Hibrit sınıflandırma, Makine öğrenmesi*

I. INTRODUCTION

Severe reactions to typically harmless substances are known as allergies. Genetic and environmental factors, including exposure to allergens and nutritional status, can have an impact on these reactions. Type 1 hypersensitivity, which is characterized by the production of particular IgE antibodies, is the most common type of allergy [1]. A subset of these allergies are food allergies, which are also IgE-mediated and usually show symptoms several hours after consumption. The quality of life for those who are impacted by these allergies is greatly diminished. Avoiding the foods that trigger an allergy is still the most effective way to manage it [2]. Still, it can be difficult to find reliable information about food allergies. Insufficient knowledge frequently compels patients and their caregivers to create personal safety plans, which exacerbates discrimination and social exclusion. Allergy reactions may pose a serious risk to life in certain situations, especially for kids, teens, or adults [3].

Children are more likely to develop food allergies than adults because of their developing immune systems and digestive systems. The majority of food allergies, including those to milk or eggs, first manifest in childhood and then disappear as the child becomes older. Childhood food allergy is a dangerous, potentially fatal illness that is known to significantly reduce patients' and their caregivers' quality of life [1]–[4]. The number of children who suffered from allergies in 2011 had doubled over the preceding ten years, according to the European Academy of Allergy and Clinical Immunology (EAACI) [5]. Food allergies were projected to affect 8% of US children in 2011 [1], with approximately 40% of those children reporting a history of serious responses. Children with food allergies to more than one food account for about 30% of cases [6]. Every three minutes, an allergic food reaction in the US takes a patient to the emergency department [7]. 20–30% of the world's population is currently affected by one or more allergies, according to the World Allergy Organization's white book of allergens, which documents the rise in the frequency of allergic disorders worldwide [8]. The data component for patients with food allergies may have associations that can be found using machine learning, which tries to teach computers how to learn and act without being explicitly taught [9]. Data mining, commonly referred to as the process of learning from experience by studying previous data, uses machine learning algorithms to collect data.

In literature, there exists a study that focuses on food allergy by using machine learning techniques. The possibility of utilizing machine learning methods to find these links is explored in [10] work. The medical laboratory Intermedia gathered the information for this investigation, which included test results from patients who had known food allergies. On this data, the apriori algorithm is used. The identified associations are put into practice in a computer program that has a data entry interface for new patients undergoing food allergy testing. In another study [11] they focused on Predictive factors for allergy at 4–6 years of age-based children. The purpose of this work was to use feature selection in machine learning to find predictors for the occurrence of parental-reported allergy at 4-6 years of age. In another research, machine learning was used to forecast how allergy challenges involving aspirin and beta-lactam allergies would turn out [12]. Applications of machine learning have also demonstrated potential in foretelling the responses of atopic dermatitis and severe asthma to biological treatments [13], [14]. Retrospective data from heated egg challenges for egg allergy were used in the study [15] to apply machine learning to the results of food allergy. Age, sex, total serum IgE, egg proteins, serum specific IgE levels, and the results of oral food challenges to the heated egg in a small cohort of 67 children with egg allergy were the training variables. The scientists observed sensitivity and specificity values for predicting heated egg challenge results of 0.51 to 0.68 and 0.66 to 0.74, respectively, along with accuracies of up to 72% using extreme gradient boosting and Support Vector Machine (SVM) models. Despite the fact that this study showed the possibility of using machine learning in real-world applications for food allergies, its generalizability was constrained by its small sample size, paucity of pertinent laboratory variables, and scarcity of clinical data. The proper classification of childhood allergies is essential for developing efficient strategies for diagnosis, treatment, and prevention. Existing

classification algorithms, however, have difficulties due to the complex and nuanced nature of childhood allergies, which is influenced by variables like demography and varied prevalence rates.

In order to accurately classify two types of asthma in preschoolers predominantly allergic asthma and non-allergic asthma—the study examined a number of machine-learning models [16]. Finding the most efficient model that could make this distinction with the fewest features was their main goal. With a high accuracy of 77.8%, the SVM with a linear kernel was found to be the most accurate model in the study for identifying preschool-aged children's asthma as either primarily allergic or non-allergic. Additionally, this model displayed a true positive rate of 0.73, a true negative rate of 0.81, and a precision of 0.81. It also achieved a ROC-AUC score of 0.79 and an F1 score of 0.81, demonstrating its efficacy in distinguishing between the two forms of asthma. Conversely, with an overall accuracy of 76.2%, which was marginally less than the SVM method but still quite good, Logistic Regression was the second-best classifier.

With an emphasis on developing a patient-specific and allergen-specific assessment for anaphylaxis risk, the study created a machine learning model using data from the Tolerance Induction Program (TIP) for anaphylactic patients [17]. This model assigns a quantitative allergen score, which improves the accessibility and accuracy of the assessment of anaphylaxis risk in children with food allergies, especially with regard to peanut allergens. The efficacy of the model was validated by applying it to a particular group of children who had experienced food allergies. The study offers data on how COVID-19 affects cardiovascular health, emphasizing the need to monitor hematological changes, recognize the occurrence of allergies, and use predictive modeling to enhance risk assessment and management techniques [18]. Another study [19] presents a systematic approach to identify amino acid subsequences (ASPs) that are more common in allergenic proteins than in non-allergenic proteins. A database of 21,154 proteins with documented allergenic reactions was assembled and examined as an example of this methodology. The ASPs found in this proof-of-concept investigation were consistent with accepted biological theories. Furthermore, the allergenicity prediction made with these discovered ASPs performed better than previous techniques. This shows that this method may be used to determine whether artificial foods and proteins are allergenic. The study [20], carried out by a different research team, concentrated on using machine-learning techniques to forecast Oral Food Challenge (OFC) results. They collected retrospective information from 1,112 patients who had 1,284 OFCs performed in total. The information included a range of clinical characteristics, including age, sex, skin prick test (SPT) results, serum-specific immunoglobulin E (IgE), and total IgE levels. Making use of this extensive dataset, the researchers built several machine-learning models. These models were created especially to forecast the results of OFCs for milk, eggs, and peanuts—three common allergens. Their method is a big step toward improving the predictability of allergic reactions in clinical settings by utilizing data-driven techniques. The problem of cow's milk allergy (CMA) was examined in a notable study [21] conducted by a different research team, with an emphasis on diagnosis and treatment. The Oral Food Challenge (OFC) is the gold standard for validating a CMA diagnosis. A thorough history evaluation and precise allergy diagnostics are the standard procedures for diagnosing CMA. The study notes that the only way to develop long-lasting milk tolerance at this time is oral immunotherapy or OIT. Another research team investigated the use of allergen multiplex assays as a precision medicine tool for patients with difficult-to-diagnose allergies [22]. These tests necessitate in-depth understanding of molecular allergy and can take a long time to interpret. The investigators postulated that the utilization of a countrywide dataset, which is equipped to facilitate allergy diagnosis through artificial intelligence (AI), could improve the care of individuals with allergies. This strategy points to a possible move in allergy care toward more data-driven, AI-enhanced diagnostic procedures.

A database was developed to examine sensitizations to 25 common aeroallergens in the Northeastern United States (zone 1) as part of a project conducted by a different research team [23]. The ImmunoCAP® in vitro assay was used to gather the data for this database. The group then performed model-based clustering using the Scikit-Learn® machine-learning library in an effort to locate allergic polysensitization clusters. After that, these clusters were examined to look for variations in the common clinical indicators of asthma seen in office environments. This strategy is an inventive application of

machine-learning to improve our knowledge of allergen sensitization patterns and how they may be related to asthma.

The purpose of the study [24] was to evaluate the predictive power of epitope-specific antibodies for an OFC given to children aged five. Of the 74 subjects, 38 had a positive OFC after five years. Using a Bead-Based Epitope Assay, IgE and IgG4 antibodies to 64 linear epitopes from Ara h 1-3 proteins were measured at different ages (4-11 months, 1 year, and 2.5 years). Furthermore, ImmunoCAP was used to measure specific IgE to peanut and component proteins. To determine the earliest time point for a reliable 5-year outcome prediction, machine-learning techniques were utilized. In order to do this, prognostic algorithms were developed using data from various age points and validated on a different cohort of ninety children.

The study [25] evaluated asthma development prediction models in children by performing a systematic review and meta-analysis in accordance with the PRISMA guidelines. The main objective was to compare models that integrate statistical techniques and risk factors with traditional approaches that involve risk factors and logical regression. On July 23, 2021, a thorough search of pertinent studies published between 2011 and 2021 was conducted using online resources such as Science Direct, PubMed, and Google Scholar. The objective was to identify and evaluate publications that were especially concerned with children's asthma prediction models, with a focus on deep learning and machine-learning techniques.

The complexities of childhood allergies, which involve the interaction of demographic factors and specific dietary allergens, are typically difficult for traditional classification algorithms to handle. The prevalence and severity of pediatric allergies may be influenced by demographic factors like age, gender, ethnicity, and socioeconomic position. Furthermore, precise classification is made more difficult by the large variety of dietary allergens and the varied symptoms displayed by affected kids. In the study [26], they used a machine-learning methodology to endotype childhood allergy rhinitis (AR), with the goal of improving diagnostic precision and enabling individualized healthcare within the domain of precision medicine. Machine learning, specifically Convolutional Neural Networks (CNNs), has emerged as a critical method for detecting skin disease allergies using dermatological images. [27] contributes to this field by using a CNN model on a dataset of 100 skin illness photographs from web sources. In terms of accuracy, precision, recall, and error rate, they compare our system to proven methodologies such as SVM, Random Forest (RF), and Naive Bayes (NB). Their findings support the suggested model's advantage in accuracy and performance.

In another study, they investigated a variety of relevant machine-learning paradigms and models [28]. Their discussion includes the complexities of model training and validation, as well as examples of machine learning's use in the domain of allergic disorders, particularly with regard to specific environmental factors. Furthermore, they are working to connect these environmental data points to the comprehensive exposome. The promise of artificial intelligence in customized medicine is highlighted, along with an investigation of methodological approaches for healthcare enhancement through sophisticated AI techniques, ultimately leading to public health improvement. In another study [20], they discuss the crucial importance of OFCs in accurately diagnosing food allergies in the study, particularly given the limitations of current clinical testing procedures. However, patient reluctance and the restricted availability of allergists in rural healthcare settings impede the widespread usage of OFCs. The findings of their study highlight the potential of ensemble learning to predict OFC outcomes and highlight crucial clinical parameters that deserve additional investigation. In the study [29], they present a pioneering chemometric approach that offers fresh insights into unraveling the allergenic properties of food proteins. Employing advanced machine-learning techniques, both supervised and unsupervised, their research endeavors to predict the allergenicity of plant-based proteins. This innovative strategy is centered around scoring descriptors and rigorously evaluating their classification efficacy. Their partitioning methodology harnesses SVM, complemented by the application of a k-nearest neighbor (KNN) classifier. Rigorous validation is achieved through a fivefold cross-validation technique, utilized not only during the variable selection phase but also in the final classifier assessment. The overarching

goal is to provide a robust and effective classification methodology for proteins, ultimately addressing the challenge of food allergies.

In our study, we developed and evaluated a novel hybrid classification method for accurately classifying childhood allergies. In order to improve our model's accuracy and interpretability, we adopt a hybrid classification strategy that combines the best features of Decision Tree and SVM classifiers. The SVM algorithm is applied first, acting as a preliminary filter in the process. In order to divide the training data into various subsets, this step entails classifying the data instances and giving each a probability score. After that, we treat each of these subsets separately with Decision Tree classifiers. Compared to using a single model on the complete dataset, this independent treatment may produce predictions that are more accurate by allowing for the exploitation of distinctive features within each subset. We call our method "hybrid" because it successfully blends the subset-specific, fine-grained classification skills of Decision Trees with the initial data filtering and classification capabilities of SVM. The limitations that are frequently present in simple algorithms which generally struggle to strike a balance between high accuracy and interpretability are intended to be addressed by this methodology. Our hybrid algorithms use SVM for preprocessing and Decision Trees for detailed model induction on the filtered data, with the goal of achieving higher accuracy without compromising the results' readability. Our study's main contribution is this creative fusion of two disparate machine-learning approaches, which offers a well-rounded answer in terms of accuracy and interpretability.

II. MATERIALS AND METHODS

The accuracy and quality of the training data are critical factors in assessing a model's performance. Recognizing this, we carefully assessed our model using reliable and accurate data to make sure it was reliable. Building on this premise, we used an innovative hybrid classification technique in our investigation. This approach was essential to both the training and evaluation of the data. We sought to improve the model's training efficiency by utilizing this hybrid classification approach and carefully choosing the best data from the dataset for classification. In this direction, machine-learning methods were implemented for classification through the dataset named allergy and compared with the proposed approach.

The dataset was carefully chosen during the training phase with a focus on finding and including data instances with a high classification success rate. This was done in an effort to speed up the model's learning process and increase its capacity to correctly categorize cases of childhood allergies such as asthma, atopic dermatitis, allergic rhinitis, and food allergies. We were able to make use of the hybrid classification algorithm's advantages and make sure that the model was trained on the most accurate and reliable data by using this strategy. We also carried out thorough testing on a second collection of high-quality data examples to assess the model's performance and generalizability. Through this procedure, we were able to evaluate the model's accuracy in categorizing childhood allergies and gain knowledge of its dependability. Our goal was to improve the model's performance and establish its reliability in the field of identifying childhood allergies, therefore we used a thorough approach to training and testing.

A. DATASET

The dataset utilized in this study is an important source for learning more about the prevalence and results of pediatric allergies' associated treatments. It provides insights into the present population of people affected by asthma, atopic dermatitis, allergic rhinitis, and food allergies through retrospective data gathered from healthcare providers [30]. The dataset is provided by The Children’s Hospital of Philadelphia. A more substantial cross-sectional cohort of 333,200 kids who began receiving primary care in this network before turning 18 and had at least 12 months of follow-up in the network. There exist fifty columns. The dataset examines data on diagnoses made by healthcare providers to ascertain the age at diagnosis, incidence, and prevalence of eczema, asthma, rhinitis, and food allergies. To more

precisely calculate the prevalence rates of asthma, the data set looks further into prescriptions for asthma-related medications. The dataset is openly available at http://dx.doi.org/10.5281/zenodo.44529, the Zenodo repository. Open-access datasets are essential because they allow the general public unfettered access to data, which promotes transparency, teamwork, and creativity in analysis and research. In Table 1, the characteristics of the data used in our study have been illustrated.

Table 1. *Sample Data from Food Allergy Dataset*

subject_ıd	bırth_year	gender_factor	ethnıcıty_factor	age_start_years	…..	allergy_type
1	2006	S1 - Female	E0 - Non-Hispanic	0.093	…..	atopıc_derm
2	1994	S1 - Female	E0 - Non-Hispanic	0.232	…..	allergıc_rhınıtıs
3	2006	S0 - Male	E1 - Hispanic	0.0108	…..	asthma

B. METHODS

Our study attempted to address the challenging task of classifying allergies in children. In order to do this, we created a novel hybrid classification approach that combines the advantages of SVM and Decision Trees. SVM is a well-liked and effective machine-learning method used for regression and classification applications. SVM seeks to maximize the distance between classes in the feature space by locating an ideal hyperplane that divides them. The fundamental idea behind SVM is to use kernel functions, which enable the detection of nonlinear correlations, to transform the input data into a higher-dimensional feature space. SVM produces a reliable and generalizable classification model by locating the hyperplane with the greatest margin. SVM has displayed exceptional performance in a range of fields, including bioinformatics, text classification, and picture recognition. SVM, however, can be delicate to the selection of hyperparameters and necessitate careful adjustment for the best outcomes [31].

Decision trees which are also used in the proposed method, are a popular and interpretable machine learning algorithm used for classification and regression tasks. They recursively partition the feature space based on a series of decision rules to create a tree-like structure. Each internal node of the tree represents a decision rule based on a specific feature, while the leaf nodes correspond to the final classification or regression outcomes. The decision tree algorithm is characterized by its ability to handle both categorical and numerical features, as well as its capacity to capture complex interactions and nonlinear relationships between the predictors. Decision trees are intuitive, as they provide clear and interpretable decision paths, making them particularly useful in domains where model transparency and explainability are important. However, decision trees are prone to overfitting and can be sensitive to small changes in the data. Various strategies, such as pruning and ensemble techniques like Random Forests, have been proposed to address these limitations and improve the performance and generalization of decision trees [32].

C. PROPOSED MODEL

Typically, basic classification algorithms induce a single model from training data; they have become known for their simplicity and ease of model interpretation. However, these algorithms frequently run into problems when trying to reach very high accuracy. Misclassification of instances is one prominent problem: a case that one model interprets incorrectly may be correctly predicted by another. The idea of

hybrid classification was born out of this observation. This methodology uses standard classification algorithms for both data preprocessing and model induction.

Misclassified instances are typically discarded as noise in conventional scenarios. This study, however, casts doubt on this notion by suggesting that these examples could nevertheless provide insightful information about the class values of other examples. The suggested hybrid classification method carefully selects training examples to create three unique models for prediction in order to fully realize this potential. Then, one of these models is used to classify each testing instance exclusively. By combining the advantages of decision tree and SVM induction classifiers, this method effectively maximizes each model's robustness and overall predictive accuracy. By overcoming the fundamental drawbacks of simple classification algorithms, this hybrid methodology seeks to provide a more sophisticated and useful framework for managing large and complicated data sets.

The algorithm is an adaptable instrument in this phase because it can handle both continuous and categorical variables, and it is strong when handling non-linear relationships. It excels at capturing the subtleties in datasets where variables interact in intricate ways, guaranteeing that the resulting model reflects not only surface-level trends but also a more profound and perceptive understanding of the dynamics of the data.

```
Algorithm: Hybrid Classification

Input: TrainingDataSet, TestingDataSet
Output: PredictedClassifications

Begin
   // Step 1: Data Preprocessing
   PreprocessedData <- PreprocessData(TrainingDataSet)

   // Step 2: Filter Training Instances
   FilteredData1, FilteredData2, FilteredData3 <- FilterTrainingInstances(PreprocessedData)

   // Step 3: Model Induction
   Model1 <- TrainDecisionTree (FilteredData1)
   Model2 <- TrainDecisionTree(FilteredData2)
   Model3 <- TrainDecisionTree(TrainingDataSet)

   // Step 4: Classification of Testing Instances
   foreach instance in TestingDataSet do
      if SuitableForModel1(instance) then
         Prediction <- ClassifyUsingModel(Model1, instance)
      else if SuitableForModel2(instance) then
         Prediction <- ClassifyUsingModel(Model2, instance)
      else
         Prediction <- ClassifyUsingModel(Model3, instance)
      end
      StorePrediction(instance, Prediction)
   end

   // Step 5: Output Predictions
   return AllStoredPredictions
End
```

This study takes a systematic approach to categorization by combining the SVM method and the Decision Tree algorithm. The primary goal is to distinguish separate classes within a dataset using a multi-step procedure that culminates in the development of the best classification model.

C. 1. Data Preparation and Division

We concentrated on the meticulous preparation and segmentation of the dataset during the first phase of our investigation. The dataset was split into two distinct portions: 40% was reserved for testing and 60% was used for training. In order to guarantee a thorough model evaluation process, this division is essential. Preventing overfitting is beneficial as it helps prevent models from performing well on training

data but poorly on fresh, untested data. We guarantee that our model learns efficiently and generalizes well to new data a crucial component of trustworthy machine-learning models by maintaining a separate testing subset.

After this deliberate partitioning of the data, the training dataset is analyzed using the Support Vector Machine (SVM) technique. The main goal of this phase is to evaluate each data instance in the training subset and ascertain whether or not it belongs to a specific class. By using a probabilistic method to determine the probability scores for each instance's classification, the SVM achieves this. These scores are essential because they offer a measurable degree of assurance in the SVM's classification judgments. This step not only involves data classification but also involves determining the degree of confidence associated with each classification, which paves the way for the subsequent stage of our model's development.

C. **2. Dataset Segmentation Based on Probability Thresholding**

Strategic dataset segmentation, informed by a calculated probability approach, is a critical component of the employed methodology. Here, a critical threshold for dataset division is ascertained by applying a machine-learning algorithm that was specifically created for classification tasks. All of the dataset's instances are given a probability score by the algorithm, which evaluates their likelihood of classification.

This probability score is then added up and divided by the total number of instances in the dataset to determine the critical threshold. The idea behind splitting the dataset into two separate parts is this computed average. A case falling into the first dataset segment is one where the probability score is at or above this average threshold. There appears to be more faith in their classification accuracy based on this categorizing. As a result of a more cautious approach to their classification, instances that fall below this average probability threshold are assigned to the second segment. By clearly separating instances according to the degree of confidence in their classification, this method guarantees a more sophisticated and efficient handling and analysis of the data. This approach improves the classification process's accuracy while also facilitating a more focused and effective data analysis.

C. **3. Decision Tree Algorithm Implementation**

In the subsequent phase of our analysis, we apply the Decision Tree algorithm to the whole dataset as well as the first and second datasets, which were previously divided. Using the Decision Tree, the datasets are analyzed during the implementation process, and conclusions are made based on the patterns found. This algorithm works especially well because it can model intricate relationships between different variables and draw clear decision boundaries. Our implementation of the Decision Tree algorithm creates a tree-like structure with nodes representing decisions based on distinct features by iteratively dividing the data according to predetermined criteria. The algorithm is guided by these choices to produce the final classification or prediction result. In order to accurately classify allergic reactions in children, we customized the Decision Tree to handle the complexities of our dataset and make sure it captures the complex relationships within the data.

C. **4. Classification Model Generation and Evaluation**

Through its application to partitioned datasets and the comprehensive dataset, the Decision Tree algorithm generates numerous categorization models. The resulting models are properly documented and thoroughly analyzed. Each model's parameters, structures, and performance indicators are meticulously maintained, allowing for an informed comparison. To ensure a multifaceted evaluation of our classification models' performance, we employed a wide range of metrics. The percentage of correct predictions was used to determine the accuracy of the model, which served as a primary indicator of its overall efficacy. In order to comprehend the accuracy of positive predictions and demonstrate the model's dependability in identifying genuine positive instances, precision was essential. Combining

recall and precision, the F1-Score provided a balanced metric that was especially useful when dealing with unbalanced datasets. Specificity assessed the model's ability to identify true negatives, making sure it wasn't unduly sensitive, while sensitivity, or recall, measured the model's capacity to accurately identify real positive cases. The model's discriminative power was evaluated using the ROC-AUC Score, where higher scores denoted better class distinction. Last but not least, the model's quality was balanced by the Matthews Correlation Coefficient (MCC), which was particularly helpful in situations where class distributions were not uniform. When combined, these metrics provided a thorough understanding of the accuracy, precision, and dependability of our model across a range of classification-related domains.

C. 5. Optimal Model Identification

A key goal of this methodological framework is to identify the best categorization model. Extensive testing of the Decision Tree models reveals the one with the highest degree of accuracy, precision, recall, or any other relevant performance parameter, indicating its superiority in capturing the underlying patterns and characteristics of the dataset.

The methodology outlined herein covers the whole workflow of this investigation, from initial data division through final model identification. The interaction of SVM-based probabilistic analysis with the Decision Tree algorithm application results in a robust and comprehensive classification scheme. The study aims to improve our understanding of classification approaches and contribute to the larger domain of data analysis and pattern recognition by revealing the intricacies of this process.

It is critical to highlight that the specific parameters, datasets, and performance indicators used in this methodology are sensitive to the context and aims of the study. The Flow chart diagram of the model exists in Figure 1.

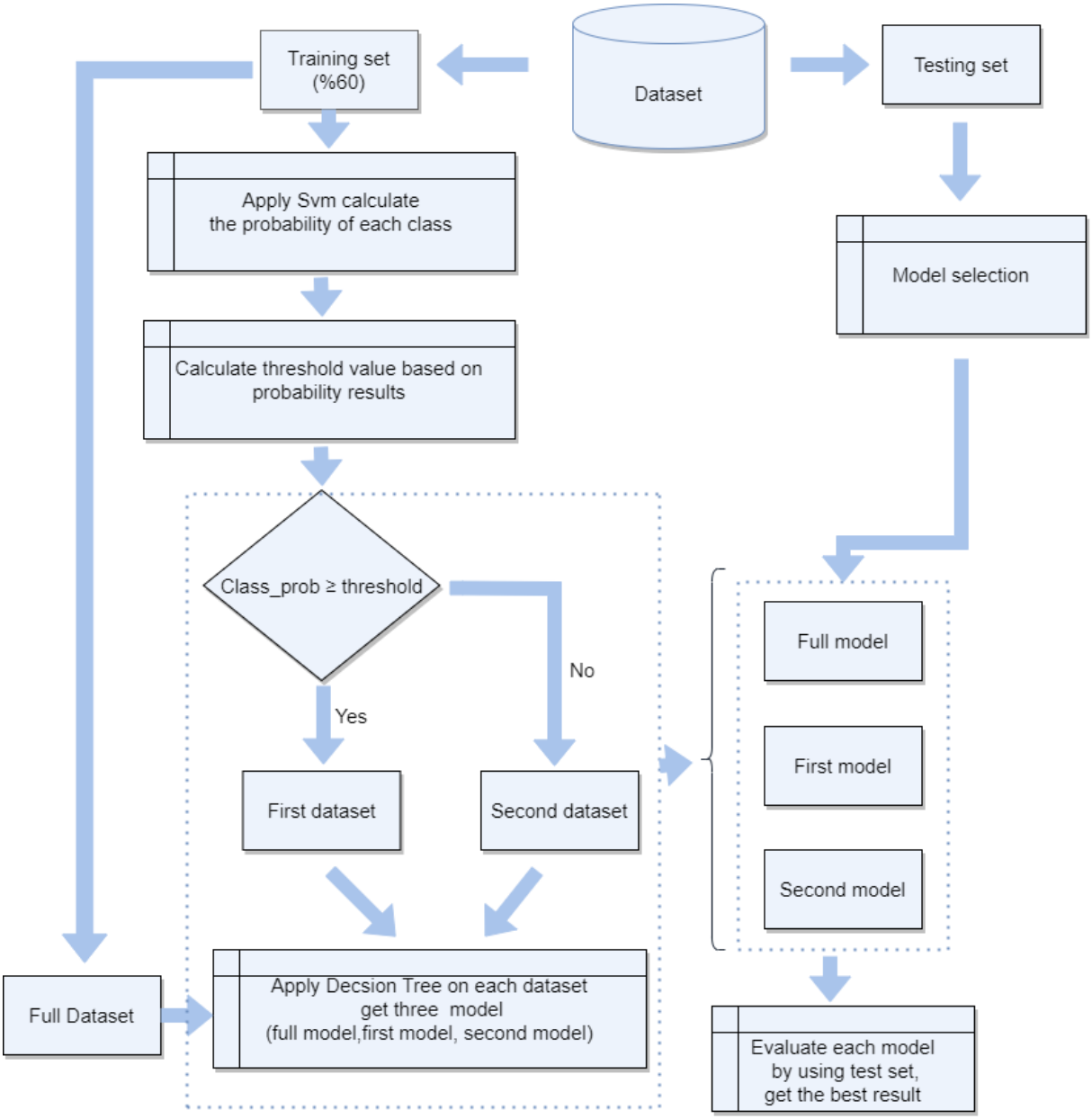

Figure 1. *Flowchart diagram of proposed method.*

In summary, this study's algorithm creates models through a multi-phase process of data handling and hybrid classification. 60% of the dataset was initially designated for training and 40% for testing. A SVM was used to process the training dataset, giving a probability score to each row of data. After that, a threshold value was determined by averaging these probability scores. The training data was then divided into two separate datasets based on this threshold. The "first dataset" contained data points with probability scores that were either equal to or higher than the threshold, while the "second dataset" contained data points with lower scores. Furthermore, the training set as a whole was maintained as the "full dataset." The first, second, and full datasets were then subjected to independent application of a Decision Tree algorithm, which produced three distinct models. The test data set was then used to apply these models for predictions. Ultimately, a thorough evaluation of the test results was conducted in order to determine how well the hybrid classification algorithm performed. This novel strategy reveals a creative way to use SVM for preliminary data slicing and Decision Tree algorithms for model building, with the goal of improving prediction accuracy while preserving interpretive simplicity.

III. RESULTS

This section presents the study's findings, focusing on the outcomes produced from the use of the SVM method for probabilistic analysis and the subsequent implementation of the Decision Tree algorithm on the segregated dataset. The primary goal was to find the best classification model for the datasets under consideration.

Upon examination of the dataset's distribution of allergy data among patients reveals a predominance of 'ATOPIC_DERM,' 'ALLERGIC_RHINITIS,' and 'ASTHMA' disorders. As a result of this finding, these three classes were chosen as the focus point of the classification problem, acting as the classification task's foundational labels Figure 2. Furthermore, when investigating the co-occurrence of allergies, it becomes evident that these same three types of allergies, namely 'ATOPIC_DERM,' 'ALLERGIC_RHINITIS,' and 'ASTHMA,' are the most commonly observed in conjunction with one another Figure 3.

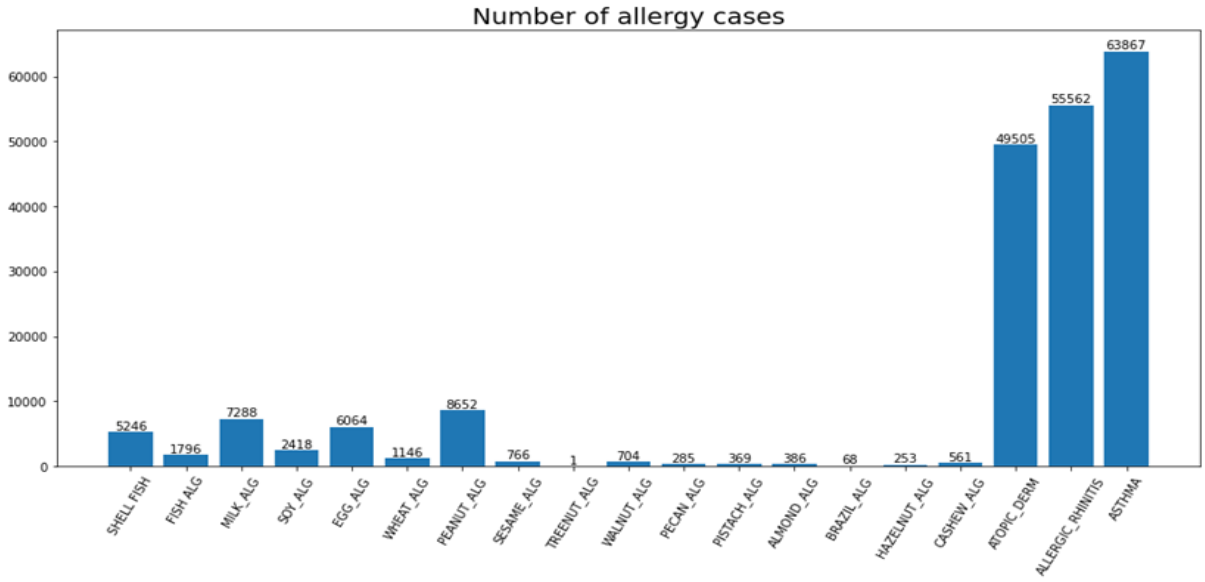

Figure 2. *Number of allergy cases*

A calculated probability-guided strategic dataset segmentation is crucial to the process. The SVM algorithm is used to generate the threshold of 0.5 for dataset segmentation, which is determined through a rigorous analytical process. In particular, the probabilities of classification likelihood for each data instance derived from the SVM output are added up, and this total is divided by the total number of instances in the dataset to determine this threshold. In essence, the dataset is divided using this method based on the average probability of classification as the critical criterion. Examples with probabilities that are at or above the 0.5 cutoff are categorized into the first dataset, indicating a higher degree of classification confidence. On the other hand, cases that are less than this probability threshold are carefully assigned to the second dataset. By efficiently distinguishing between examples with greater and lower classification certainty, this binary segmentation technique ensures a more focused and sophisticated approach in the dataset division. Following this analysis, the first dataset was made up of 88552 data rows with probabilities more than 0.5, whereas the second dataset was made up of 80082 data rows with probabilities less than 0.5.

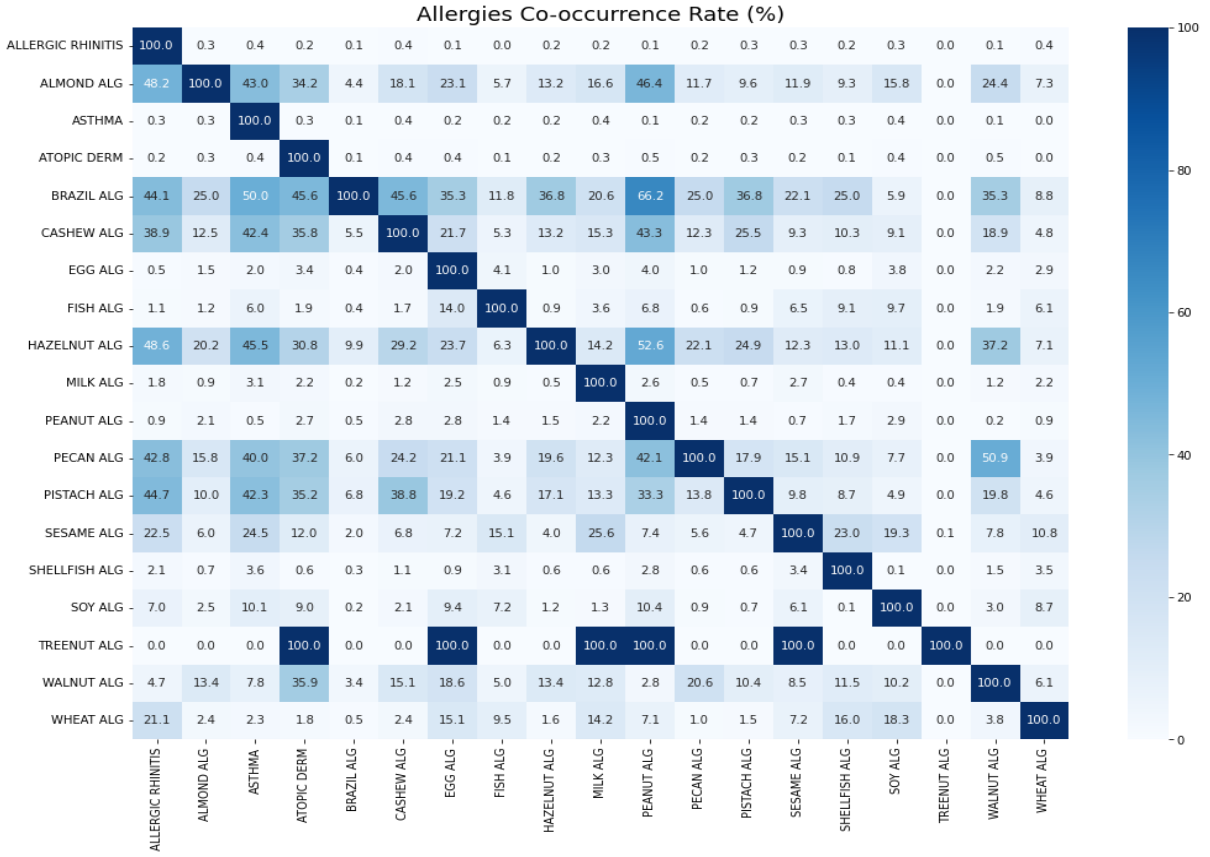

Figure 3. *Allergies Co-occurrence Rate*

In Table 2, the results for the allergy dataset are presented across various models, revealing significant differences in performance metrics.

Table 2. *Result of the allergy dataset.*

Data Slices	Accuracy	Precision	F1-Score	Sensitivity (Recall)	Specificity	ROC-AUC Score	MCC
First Part	0.801	0.812	0.766	0.831	0.749	0.788	0.678
Second Part	0.713	0.719	0.711	0.712	0.714	0.713	0.690
Orginal With Decsion Tree	0.682	0.671	0.670	0.825	0.792	0.811	0.713
Orginal With SVM	0.691	0.683	0.679	0.688	0.676	0.679	0.611
Orginal With Naive Bayes	0.685	0.675	0.670	0.780	0.800	0.710	0.632
Orginal With Random Forest	0.695	0.690	0.688	0.700	0.810	0.720	0.628
Orginal With Logistic Regression	0.680	0.672	0.675	0.690	0.795	0.705	0.622
Proposed Hybrid Model	**0.906**	**0.901**	**0.893**	**0.917**	**0.895**	**0.932**	**0.874**

First Part of Data: This model performs well, correctly predicting 80.1% of the instances with an accuracy of 0.801. With a precision of 0.812, it indicates that a high percentage of all positive predictions were true positives. Recall and precision are balanced by the F1-score of 0.766, which is a respectably high score. The model's sensitivity (recall) of 0.831 suggests that it performs a good job of identifying true positives. With a specificity of 0.749, true negatives can be identified with a moderate degree of accuracy. The model performs well, as evidenced by the ROC-AUC score of 0.788 and MCC of 0.678, the latter of which shows a balanced measure taking into account all four confusion matrix categories.

Second Part of the Data: The performance of this model has declined noticeably on the majority of metrics. Lower efficacy in accurate predictions and positive identifications is indicated by an accuracy of 0.713 and a precision of 0.719. Moderate performance is indicated by the F1-score of 0.711, sensitivity of 0.712, and specificity of 0.714, which are all above average. Even though they still show fair performance, the ROC-AUC score of 0.713 and the MCC of 0.690 are lower than those of other models.

Original Data(Whole Training Data) with Decision Tree and SVM: In comparison to the original model, both of these variants perform less well. A moderate degree of predictive accuracy is indicated by the Decision Tree variant's accuracy of 0.682, precision of 0.671, and F1-score of 0.670. While the ROC-AUC score of 0.811 and specificity of 0.792 are lower than the sensitivity of 0.825, respectively, they indicate that the former is less effective in identifying true negatives. MCC is reasonable at 0.713. The SVM variant performs modestly overall but in a balanced manner, with slightly better metrics (F1-score of 0.679, accuracy of 0.691, precision of 0.683), but lower sensitivity (0.688) and specificity (0.676). The MCC of 0.611 and the ROC-AUC score of 0.679 both attest to this moderate efficacy.

The hybrid model that has been suggested performs noticeably better than the others, with an accuracy of 0.906. This indicates that the model has good predictive power. A high degree of accuracy in positive predictions and a strong balance between precision and recall are indicated by the precision of 0.901 and an F1-score of 0.893. Its high sensitivity of 0.917 and specificity of 0.895 shows that it has the remarkable capacity to distinguish between true positives and true negatives. The remarkable MCC of 0.874 and ROC-AUC score of 0.932 indicate excellent overall performance and reliability.

Three of the most popular classification algorithms in the literature Random Forests, Logistic Regression, and Naive Bayes were compared to our model in this study. The performance metrics from the original dataset that is, the complete training dataset were the basis for the comparison. Even though they were efficient, Random Forest, Logistic Regression, and Naive Bayes performed worse than our model. For example, the Random Forest model showed good but not exceptional classification capabilities with moderate accuracy (0.695) and F1-Score (0.688). With accuracy scores of 0.680 and 0.685, respectively, Logistic Regression and Naive Bayes demonstrated comparable trends but fell short of the hybrid model.

The Proposed Hybrid Model performs noticeably better than the other models, indicating its superior predictive power and dependability when it comes to categorizing the allergy dataset.

IV. DISCUSSION

A thorough examination of allergic data within the dataset gave useful insights into the prevalence and co-occurrence of various allergic disorders in this study. The prevalence of 'ATOPIC_DERM,' 'ALLERGIC_RHINITIS,' and 'ASTHMA' among patients emphasizes the importance of these specific allergens as key classification subjects. As a result, these three classes were chosen as the primary labels for the classification problem, allowing for a more focused and informative examination.

In this study, instance filtering is used to create hybrid classification algorithms in a novel way. In contrast to simple classification algorithms, which yield one model, our approach yields three different

models per hybrid algorithm. One of these models is the only one that can classify a new instance, so the interpretation of the predictions is still as straightforward as with simple algorithms. Thus, our hybrid algorithms not only improve the predictive power and accuracy but also maintain the simple interpretability feature of simple classification models.

The use of the Decision Tree algorithm as a classification tool enabled the creation of numerous models, each representing a different component of the dataset's complexity and categorization limits. A variety of models were developed by carefully tweaking hyperparameters and in-depth study of the resulting tree structures, with varied trade-offs between precision, recall, and overall accuracy.

The identification of the best model, namely the proposed hybrid model, demonstrated the effectiveness of the presented method. This model obtained %90 accuracy on the dataset, with the precision of 0.901 and an F1-score of 0.893. Its high sensitivity of 0.917 and specificity of 0.895 demonstrate a harmonic balance between classification performance and prediction capacity. Importantly, this approach not only advances our understanding of allergic illness classification, but it also has practical applications in medical diagnostics, treatment recommendations, and public health actions.

It is critical to recognize the study's limitations. The use of specific classes and the Decision Tree technique may have consequences for generalization to other datasets and algorithms. Furthermore, while this study provides useful insights into the classification of allergy diseases, future research could look into the incorporation of other features, different algorithms, or ensemble techniques to broaden the scope of classification accuracy.

Important numerical results highlighted the empirical benefits and outcomes. In the best model, the hybrid classification algorithms, creatively employed instance filtering. The numerical result not only confirms the effectiveness of the algorithm but also marks a significant advancement in the analysis of medical data, particularly in the area of allergic disease classification. These numerical results have significant practical implications that improve public health decisions and medical diagnostics. Future research is made possible by the study's methodology, which skillfully balanced classification performance and prediction capacity. These impressive numerical results highlight the transformative potential of incorporating cutting-edge machine-learning techniques into medical science. These findings demonstrate the critical role that quantitative analysis plays in improving patient outcomes and healthcare.

The comparative analysis of various studies employing machine-learning methods in respiratory and allergic conditions highlights the diversity and effectiveness of different approaches (Table 3). While studies like Azam M.A. (2018) and Tinschert P. (2020) demonstrate moderate success with traditional models like SVM and decision trees, they are somewhat limited by smaller sample sizes or lower performance metrics. On the other hand, studies like Tenero L. and Adhi Pramono R.X. (2019) showcase the potential of more sophisticated methods like PCA and logistic regression, achieving higher sensitivity and specificity. Notably, our Proposed Hybrid Model, applied to a substantial dataset of 333,200 children, outperforms others in key metrics like accuracy and precision. This suggests that advanced, hybrid approaches can significantly enhance the capability of machine-learning models in accurately diagnosing and understanding complex medical conditions, offering promising avenues for future research and clinical application.

Table 3. *Comparative Analysis of Machine-Learning Methods in Respiratory and Allergic Condition Studies.*

Study	Participants [DataSource]	ML Methods	Performance
Azam M.A,2018 [33]	50 individuals (age notspecified) with COPD, asthma, bronchitis, andpneumonia	Bag-of-Features, SVM	F1-score =75%, accuracy rate = 75.21(complete cycle

Table 3 (cont). *Comparative Analysis of Machine-Learning Methods in Respiratory and Allergic Condition Studies.*

Tinschert P,2020 [34]	79 adults	Mixed-effects regressions, decision trees	56% <balanced accuracy <70%
Tenero L. [35]	38 children (age 6-16): asthma = 28, control = 10	PCA, penalized logistic model	Sensitivity =79%, specificity= 84%, cross-validated AUC= 80%
Adhi Pramono R.X, 2019 [36]	Unknown individuals from multiple repositories	Logistic regression	Sensitivity =86.78%, specificity =99.42%, F1-score =88.74%
Purnomo A.T, 2021 [37]	Unknown individuals	XGBoost	Precision >80%, sensitivity> 70%, F1-score > 75%(for all classes, MFCC feature extraction
Zhang O,2020 [20]	2010 individuals (age >16) with severe and persistent asthma	Recursive feature elimination, PCA, random under-sampling, random over-sampling, SMOTE, logistic regression, Naive Bayes, decision tree, perceptron	Sensitivity =90%, specificity= 83%, AUC =85%
Proposed Hybrid Model	**333,200 kids with 'atopıc_derm,' 'allergıc_rhınıtıs,' and 'asthma,'**	**Proposed Hybrid Classfication Method**	**Accuracy : %91, Precision: %90, F1-Score: %89, Sensitivity:% 0.89**

IV. CONCLUSION

In this study, instance filtering is used to create hybrid classification algorithms in a novel way. In contrast to simple classification algorithms, which yield one model, our approach yields three different models per hybrid algorithm. One of these models is the only one that can classify a new instance, so the interpretation of the predictions is still as straightforward as with simple algorithms. Thus, our hybrid algorithms not only improve the predictive power and accuracy but also maintain the simple interpretability feature of simple classification models. The combination of thorough data analysis, focused class selection, and the capability of the Decision Tree algorithm has resulted in a strong framework for allergic disease categorization. The findings given in this paper add to the larger field of medical data analysis and categorization approaches. This study not only provides a more nuanced view of allergy disorders, but it also opens the door to future research targeted at improving our ability to

grasp, diagnose, and treat a wide range of medical conditions using modern machine-learning techniques.

V. REFERENCES

[1] R. S. Gupta *et al.*, "The public health impact of parent-reported childhood food allergies in the United States," *Pediatrics*, vol. 142, no. 6, p. e20181235, 2018.

[2] S. M. Jones and A. W. Burks, "Food allergy," *New England Journal of Medicine*, vol. 377, no. 12, pp. 1168–1176, 2017.

[3] A. Elghoudi and H. Narchi, "Food allergy in children—the current status and the way forward," *World Journal of Clinical Pediatrics*, vol. 11, no. 3, p. 253, 2022.

[4] C. Westwell-Roper *et al.*, "Food-allergy-specific anxiety and distress in parents of children with food allergy: A systematic review," *Pediatric Allergy and Immunology*, vol. 33, no. 1, p. e13695, 2022.

[5] E. Jensen-Jarolim *et al.*, "State-of-the-art in marketed adjuvants and formulations in allergen immunotherapy: a position paper of the European Academy of Allergy and Clinical Immunology (EAACI)," *Allergy*, vol. 75, no. 4, pp. 746–760, 2020.

[6] P. Bégin *et al.*, "CSACI guidelines for the ethical, evidence-based and patient-oriented clinical practice of oral immunotherapy in IgE-mediated food allergy," *Allergy, Asthma & Clinical Immunology*, vol. 16, pp. 1–45, 2020.

[7] S. Clark, J. Espinola, S. A. Rudders, A. Banerji, and C. A. Camargo, "Frequency of US emergency department visits for food-related acute allergic reactions," *Journal of allergy and clinical immunology*, vol. 127, no. 3, pp. 682–683, 2011.

[8] R. Pawankar *et al.*, "Asia Pacific Association of Allergy Asthma and Clinical Immunology White Paper 2020 on climate change, air pollution, and biodiversity in Asia-Pacific and impact on allergic diseases," *Asia Pacific Allergy*, vol. 10, no. 1, 2020.

[9] C. J. Haug and J. M. Drazen, "Artificial intelligence and machine learning in clinical medicine, 2023," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1201–1208, 2023.

[10] A. Ktona, A. Mitre, D. Shehu, and D. Xhaja, "Support Allergic Patients, using Models Found by Machine Learning Algorithms, to Improve their Quality of Life.," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 10, no. 4, pp. 512–517, 2022.

[11] K. Kamphorst, A. Lopez-Rincon, A. M. Vlieger, J. Garssen, E. van't Riet, and R. M. van Elburg, "Predictive factors for allergy at 4–6 years of age based on machine learning: A pilot study," *PharmaNutrition*, vol. 23, p. 100326, 2023.

[12] E. M. Moreno *et al.*, "Usefulness of an artificial neural network in the prediction of β-lactam allergy," *The Journal of Allergy and Clinical Immunology: In Practice*, vol. 8, no. 9, pp. 2974–2982, 2020.

[13] J. J. Wu *et al.*, "Predictors of nonresponse to dupilumab in patients with atopic dermatitis: a machine learning analysis," *Annals of Allergy, Asthma & Immunology*, vol. 129, no. 3, pp. 354–359, 2022.

[14] D. Di Bona, F. Spataro, P. Carlucci, G. Paoletti, and G. W. Canonica, "Severe asthma and personalized approach in the choice of biologic," *Current Opinion in Allergy and Clinical Immunology*, vol. 22, no. 4, pp. 268–275, 2022.

[15] Y. Kuniyoshi, H. Tokutake, N. Takahashi, A. Kamura, S. Yasuda, and M. Tashiro, "Machine learning approach and oral food challenge with heated egg," *Pediatric Allergy and Immunology*, vol. 32, no. 4, pp. 776–778, 2021.

[16] P. Bhardwaj, A. Tyagi, S. Tyagi, J. Antão, and Q. Deng, "Machine learning model for classification of predominantly allergic and non-allergic asthma among preschool children with asthma hospitalization," *Journal of Asthma*, vol. 60, no. 3, pp. 487–495, 2023.

[17] I. S. Randhawa, K. Groshenkov, and G. Sigalov, "Food anaphylaxis diagnostic marker compilation in machine learning design and validation," *Plos one*, vol. 18, no. 4, p. e0283141, 2023.

[18] M. G. Yousif, F. G. Al-Amran, A. M. Sadeq, and N. G. Yousif, "The Impact of COVID-19 on Cardiovascular Health: Insights from Hematological Changes, Allergy Prevalence, and Predictive Modeling," *Medical Advances and Innovations Journal*, vol. 1, no. 1, p. 10, 2023.

[19] K. Goto *et al.*, "Novel machine learning method allerStat identifies statistically significant allergen-specific patterns in protein sequences," *Journal of Biological Chemistry*, vol. 299, no. 6, 2023.

[20] J. Zhang *et al.*, "Prediction of oral food challenge outcomes via ensemble learning," *Informatics in Medicine Unlocked*, vol. 36, p. 101142, 2023.

[21] M. A. Tosca, R. Olcese, C. Trincianti, M. Naso, I. Schiavetti, and G. Ciprandi, "Children with cow's milk allergy: prediction of oral immunotherapy response in clinical practice," *Allergo Journal International*, pp. 1–2, 2023.

[22] G. Martinroche *et al.*, "Creating a French Dataset for artificial intelligence-assisted allergy diagnosis using semantic attributes and allergen multiplex technology," *Journal of Allergy and Clinical Immunology*, vol. 151, no. 2, p. AB318, 2023.

[23] B. J. Patchett *et al.*, "Allergic Polysensitization Clusters: Newly Recognized Severity Marker in Urban Asthmatic Adults," *International Archives of Allergy and Immunology*, vol. 184, no. 3, pp. 261–272, 2023.

[24] S. Grinek *et al.*, "Epitope-specific IgE at 1 year of age can predict peanut allergy status at 5 years," *International Archives of Allergy and Immunology*, vol. 184, no. 3, pp. 273–278, 2023.

[25] R. H. Ekpo, V. C. Osamor, A. A. Azeta, E. Ikeakanam, and B. O. Amos, "Machine learning classification approach for asthma prediction models in children," *Health and Technology*, vol. 13, no. 1, pp. 1–10, 2023.

[26] V. Malizia *et al.*, "Endotyping allergic rhinitis in children: A machine learning approach," *Pediatric Allergy and Immunology*, vol. 33, pp. 18–21, 2022.

[27] V. R. Allugunti, "A machine learning model for skin disease classification using convolution neural network," *International Journal of Computing, Programming and Database Management*, vol. 3, no. 1, pp. 141–147, 2022.

[28] M. H. Shamji *et al.*, "EAACI guidelines on environmental science in allergic diseases and asthma–Leveraging artificial intelligence and machine learning to develop a causality model in exposomics," *Allergy*, 2023.

[29] M. Nedyalkova, M. Vasighi, A. Azmoon, L. Naneva, and V. Simeonov, "Sequence-Based Prediction of Plant Allergenic Proteins: Machine Learning Classification Approach," *ACS omega*, vol. 8, no. 4, pp. 3698–3704, 2023.

[30] D. A. Hill, R. W. Grundmeier, G. Ram, and J. M. Spergel, "The epidemiologic characteristics of healthcare provider-diagnosed eczema, asthma, allergic rhinitis, and food allergy in children: a retrospective cohort study," *BMC pediatrics*, vol. 16, no. 1, pp. 1–8, 2016.

[31] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Computers & Operations Research*, vol. 152, p. 106131, 2023.

[32] X. Han, X. Zhu, W. Pedrycz, and Z. Li, "A three-way classification with fuzzy decision trees," *Applied Soft Computing*, vol. 132, p. 109788, 2023.

[33] M. A. Azam, A. Shahzadi, A. Khalid, S. M. Anwar, and U. Naeem, "Smartphone based human breath analysis from respiratory sounds," presented at the 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018, pp. 445–448.

[34] P. Tinschert *et al.*, "Nocturnal cough and sleep quality to assess asthma control and predict attacks," *Journal of asthma and allergy*, pp. 669–678, 2020.

[35] L. Tenero, M. Sandri, M. Piazza, G. Paiola, M. Zaffanello, and G. Piacentini, "Electronic nose in discrimination of children with uncontrolled asthma," *Journal of Breath Research*, vol. 14, no. 4, p. 046003, 2020.

[36] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "Automatic cough detection in acoustic signal using spectral features," presented at the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 7153–7156.

[37] A. T. Purnomo, D.-B. Lin, T. Adiprabowo, and W. F. Hendria, "Non-contact monitoring and classification of breathing pattern for the supervision of people infected by COVID-19," *Sensors*, vol. 21, no. 9, p. 3172, 2021.