A machine learning approach for predicting familial and sporadic disease cases based on clinical symptoms: introduction of a new dataset

Klinik belirtilere dayalı ailesel ve sporadik hastalık vakalarını tahmin etmek için bir makine öğrenimi yaklaşımı: yeni bir veri kümesinin tanıtımı

Parisa SHARAFİ[1] (ID), Hilal ARSLAN[2] (ID), Sibel ERSOY EVANS[3] (ID), Ali VARAN[4] (ID), Şükriye AYTER[1] (ID)

ABSTRACT

Objective: Neurofibromatosis type 1 (NF1) is a common yet complex neurogenetic disorder characterized by a highly variable clinical presentation, influenced by both genetic and environmental factors. While its genetic basis is well understood, the variability in symptoms among patients presents significant challenges for diagnosis and management. This study focuses on examining the differences in clinical features between sporadic and familial NF1 cases. Additionally, it evaluates the potential of machine learning techniques to predict sporadic NF1 cases based on clinical symptoms, offering insights into how computational approaches can complement traditional diagnostic methods.

Methods: A retrospective analysis was conducted on the medical records of 241 NF1 patients, including 121 sporadic and 120 familial cases. The frequency of various clinical features, such as Lisch nodules, pseudoarthrosis, and hypertension, was compared between the groups. analysis of variance (ANOVA) was used to identify the most important features distinguishing sporadic cases from familial ones. Furthermore, multiple machine

ÖZET

Amaç: Nörofibromatozis tip 1 (NF1), hem genetik hem de çevresel faktörlerden etkilenen, oldukça değişken bir klinik sunumla karakterize, yaygın ancak karmaşık bir nörogenetik bozukluktur. Genetik temeli iyi anlaşılmış olsa da, hastalar arasındaki semptomların değişkenliği tanı ve yönetim için önemli zorluklar ortaya koymaktadır. Bu çalışma, sporadik ve ailesel NF1 vakaları arasındaki klinik özelliklerdeki farklılıkları incelemeyi amaçlamıştır. Ayrıca, makine öğrenimi tekniklerinin klinik semptomlara dayalı olarak sporadik NF1 vakalarını tahmin etme potansiyelini değerlendirerek, hesaplamalı yaklaşımların geleneksel tanı yöntemlerini nasıl tamamlayabileceğine dair içgörüler sunulmuştur.

Yöntem: 121 sporadik ve 120 ailesel vaka dahil olmak üzere 241 NF1 hastasının tıbbi kayıtları üzerinde retrospektif bir analiz yapılmıştır. Lisch nodülleri, psödoartroz ve hipertansiyon gibi çeşitli klinik özelliklerin sıklığı gruplar arasında karşılaştırılmıştır. Sporadik vakaları ailesel olanlardan ayıran en önemli özellikleri belirlemek için varyans analizi (ANOVA) kullanılmıştır. Ayrıca, belirlenen özelliklere dayanarak

[1]TOBB University of Economics and Technology, Faculty of Medicine, Department of Medical Biology, Ankara, Türkiye
[2]Yıldırım Beyazıt University, Faculty of Engineering and Natural Sciences, Department of Software Engineering, Ankara, Türkiye
[3]Hacettepe University, School of Medicine, Department of Dermatology, Ankara, Türkiye
[4]Hacettepe University, School of Medicine, Department of Pediatrics, Pediatric Oncology, Ankara, Türkiye

İletişim / Corresponding Author : Parisa SHARAFİ
TOBB Ekonomi ve Teknoloji Üniversitesi, Sogutozu Cd. No: 43, Ankara - Türkiye
E-posta / E-mail : parisasharafi@gmail.com

Geliş Tarihi / Received : 14.10.2024
Kabul Tarihi / Accepted : 29.12.2024

DOI ID : 10.5505/TurkHijyen.2025.06337

Sharafi P, Arslan H, Ersoy Evans S, Varan A, Ayter Ş. A machine learning approach for predicting familial and sporadic disease cases based on clinical symptoms: Introduction of a new dataset. Turk Hij Den Biyol Derg, 2025; 82(1): 99 - 106

learning algorithms, including k-nearest neighbors, artificial neural networks, support vector machines, decision trees, and XGBoost, were employed to predict sporadic cases based on the identified features.

Results: Among the machine learning models tested, the XGBoost algorithm demonstrated the highest predictive accuracy at 62.86%, indicating moderate reliability in identifying sporadic cases. Despite this limitation, the analysis revealed significant differences in clinical manifestations between the two groups. These differences suggest that shared genetic modifiers may play a critical role in shaping the observed genotype-phenotype relationship in NF1.

Conclusion: This study represents the first detailed comparison of a broad spectrum of clinical symptoms between sporadic and familial NF1 cases. While machine learning models showed only moderate success in prediction, the findings provide valuable insights into the phenotypic variability of NF1 and underscore the importance of larger, more diverse datasets for improving predictive accuracy. These results hold significant potential for guiding personalized diagnostic and therapeutic strategies for NF1 patients.

Key Words: Neurofibromatosis type 1, sporadic cases, familial cases, machine learning

sporadik vakaları tahmin etmek için k-en yakın komşular, yapay sinir ağları, destek vektör makineleri, karar ağaçları ve XGBoost dahil olmak üzere çoklu makine öğrenimi algoritmaları kullanılmıştır.

Bulgular: Test edilen makine öğrenimi modelleri arasında XGBoost algoritması %62,86 ile en yüksek tahmin doğruluğunu göstermiş ve sporadik vakaların belirlenmesinde orta düzeyde güvenilirliğe işaret etmiştir. Bu sınırlamaya rağmen, analiz iki grup arasında klinik belirtiler açısından önemli farklılıklar olduğunu ortaya koymuştur. Bu farklılıklar, paylaşılan genetik değiştiricilerin NF1'de gözlenen genotip-fenotip ilişkisini şekillendirmede kritik bir rol oynayabileceğini düşündürmektedir.

Sonuç: Bu çalışma, sporadik ve ailesel NF1 vakaları arasında geniş bir klinik semptom spektrumunun ilk ayrıntılı karşılaştırmasını temsil etmektedir. Makine öğrenimi modelleri tahminde yalnızca orta düzeyde başarı gösterirken, bulgular NF1'in fenotipik değişkenliği hakkında değerli bilgiler sağlamakta ve tahmin doğruluğunu artırmak için daha büyük, daha çeşitli veri kümelerinin öneminin altını çizmektedir. Bu sonuçlar, NF1 hastaları için kişiselleştirilmiş tanı ve tedavi stratejilerine rehberlik etme konusunda önemli bir potansiyele sahiptir.

Anahtar Kelimeler: Nörofibromatozis tip 1, sporadik vakalar, ailesel vakalar, makine öğrenmesi

INTRODUCTION

Neurofibromatosis type 1 (NF1; OMIM 162200) is the most common neurogenetic disorder, affecting approximately 1 in 3,500 individuals worldwide. It is caused by mutations in the NF1 gene located on chromosome 17q11.2 (1). This gene spans 350 kb of genomic DNA and contains 60 exons, coding for neurofibromin, a cytoplasmic protein that negatively regulates RAS proteins (2). The loss of neurofibromin leads to the activation of the RAS cascade, resulting in increased cell proliferation, categorizing NF1 as a tumor suppressor gene. The NF1 gene also has one of the highest germline mutation rates reported in humans.

About 50% of NF1 patients have sporadic cases, where de novo germline mutations occur without a family history of the disease (3-5). The clinical presentation of NF1 is highly variable and includes symptoms such as café-au-lait spots (CALS), cutaneous neurofibroma (cNF), plexiform neurofibroma (pNF), freckling, peripheral nerve sheath tumors (MPNST), malignancies, and juvenile myelomonocytic leukemia (6). Despite its autosomal-dominant inheritance, the expression of clinical symptoms is highly unpredictable, with no clear genotype-phenotype correlations, making diagnosis and management challenging (7). Modifier genes and environmental factors, such as nutrition and lifestyle, may contribute to the observed clinical variability (8).

TOBB University Faculty of Medicine has been at the forefront of NF1 research, establishing a multidisciplinary "NF Study Group" in 2003. This group includes physicians from various clinical departments, as well as researchers from the basic sciences, focusing on the molecular aspects of NF1. When examining the incidence of clinical phenotypes such as IQ expression and learning disability, differences between familial and sporadic cases have been observed (9, 10). Through the evaluation of a national NF1 database, encompassing 241 probands (121 sporadic and 120 familial cases), we aimed to explore the differences in clinical symptom expression between sporadic and familial cases.

In this study, we introduce a novel machine learning-based approach to predict sporadic NF1 cases based on clinical symptoms. Utilizing a range of algorithms, including k-nearest neighbors, artificial neural networks, support vector machines, decision trees, and gradient boosting, we sought to classify patients as sporadic or familial. The coexistence of NF1 symptoms has been previously published (11). This is the first study to employ machine learning in the prediction of sporadic NF1 cases, leveraging detailed clinical data from a well-characterized national cohort.

MATERIAL and METHOD

Dataset

In this study a national NF1 database with 241 probands (121 sporadic and 120 familial cases) was evaluated. DNAs from patients were used for mutation analysis and various number of known and novel mutations were characterized in our study and published elsewhere (12). The National Institutes of Health NF1 (NIH-NF1) clinical criteria were used for diagnosis of patients with NF1 (1, 13, 14). All participants had filled up the informed consent through a questionnaire with detailed clinical information of patients and completed by a dermatologist, a pediatric neurologist, or a clinical geneticist prospectively. The questionnaire encompassed clinical features of NF1 such as tumors and other neurological problems. The guidance of the Declaration of Helsinki were turned to account for clinical and genetic analyses and the questionnaire approved by the ethical committee of Hacettepe University. Clinical data were saved in our in-house developed database. The data that support the evidence of this study are available from the corresponding author upon reasonable request. You can access the dataset at "UCI Machine Learning Repository, Neurofibromatosis Type 1; Clinical Symptoms of Familial and Sporadic Cases."

Machine Learning Methods

We perform k-nearest neighbors, artificial neural networks, support vector machines, decision trees and gradient boosting techniques to predict whether a person has a sporadic or not based on the symptoms. In the next section, we describe these algorithms.

K-Nearest Neighbors (KNN)

The KNN (15) is a supervised machine learning method of the classification of unassigned data to the most ideal class according to the distance with other data in training set. The algorithm classifies new data points based on the comparison of features with the labeled data points in the training set. Since KNN does not need any training data points for model generation, it is a lazy algorithm. It uses all training data in the testing phase.

Artificial Neural Networks (ANN)

Artificial Neural Networks (16) is a machine learning model that makes decisions in a manner similar to the human brain. Every neural network consists of layers of nodes, or artificial neurons, an input layer, single or multiple hidden layers, and an output layer. Each node connects to others, and is associated with a particular weight and threshold. If the output of any individual node crosses the specified threshold value, then the node is activated and it transmits the data to the next layer in the network. Neural networks learn and improve their accuracy over time by training the data.

Support Vector Machines (SVM)

The SVM (17) used in machine learning is an effective learning algorithm for classification and regression problems. Its main purpose is to classify data points by creating the best hyperplane between two classes. In this process, support vectors are important points that maximize the marginal gap between classes. SVM can also work successfully in high dimensional spaces by using kernel functions. It exhibits strong performance in classification and regression tasks.

Decision Tree (DT)

The DT (18) used in machine learning is a modelling technique used to analyze data sets and extract patterns. This model consists of decision nodes and branches connecting these nodes. Each node represents a particular feature, and the branches represent the values of the features. The tree divides the data set into subgroups and makes predictions for each subgroup. The learning process takes place by routing the instances to the branches of the tree according to their feature values and making decisions at each node. DTs are used in classification and regression tasks; the parameters affect the performance of the model. In machine learning, this model is an effective tool for solving data analytics and classification problems.

Gradient Boosting (XGBoot)

XGBoost is an enhanced variant of gradient boosting, intertwining gradient descent with boosting techniques. By integrating multiple weak base classifiers into a robust ensemble, the XGBoost algorithm achieves superior classification capabilities. Unlike conventional boosting algorithms, which balance positive and negative samples, XGBoost ensures global convergence by tracking the direction of a negative gradient. This method is elucidated in (19). Furthermore, XGBoost introduces advanced regularization techniques such as L1 regularization (Lasso) and L2 regularization (Ridge), refining model generalization capabilities. Through these regularization methods, XGBoost mitigates overfitting more effectively compared to traditional gradient boosting algorithms, thereby enhancing model performance and robustness.

The study was approved by the Institutional Ethics Committee of Hacettepe University (Date: 08.11.2016 and Number: 16969557-1117).

RESULTS

The medical records of 241 patients diagnosed with NF1, comprising 121 sporadic and 120 familial cases, were retrospectively analyzed. The occurrence rates of various phenotypes across sporadic, familial, and total NF1 patients were assessed. This study also aimed to evaluate the effectiveness of several machine learning techniques in predicting sporadic cases based on clinical symptoms. In the following, we present the experimental setup and the performance metrics used to assess the machine learning models.

Performance Measures

All computational analyses were conducted on a system powered by an Intel Core i7 processor (2.6 GHz) with 16 GB of RAM, running Ubuntu 18.04.03 LTS. The machine learning algorithms were implemented using the Scikit-learn library, a comprehensive open-source toolkit in Python. The dataset was divided into training and validation sets with a 70:30 split, whereby 70% of the data was utilized for training the models and the remaining 30% for validation.

Evaluation Criteria

The machine learning models were evaluated based on four key metrics: precision, recall, F-measure, and accuracy, as detailed in Table 1.

Evaluation of Machine Learning Techniques for Predicting Sporadic Cases Using the Newly Created Dataset

Various machine learning techniques were employed to identify sporadic cases of NF1. The performance of these models is summarized in the subsequent sections. Initially, we assessed the importance of different features using the analysis of variance (ANOVA) method, with the results presented in Table 2. According to these results, significant features for detecting sporadic cases include Lisch nodules, pseudoarthrosis, hypertension, myelodysplastic syndrome, leukemia, and bone dysplasia. Conversely, scoliosis, maternal age, rhabdomyoma, and mental retardation were among the least important features.

Table 1. Metrics for evaluating machine learning model performance

Measure	Formula
Precision	(TP) / (TP+FP)
Recall	(TP) / (TP+FN)
F-measure	(2 x Precision x Recall) / (Precision+ Recall)
Accuracy (Acc)	(TP+TN) / (FN+TP+TN+FP)

Table 2. Feature importance based on ANOVA F-score

Feature Name	F-Score	Feature Name	F-Score	Feature Name	F-Score
Lisch Nodule	1.95	Ganglioblastoma	1.05	cNF	0.67
Pseudoarthrosis	1.67	MPNST	0.96	CALS	0.54
Hypertension	1.48	Astrocytoma	0.95	pNF	0.41
Myelodysplastic Syndrome	1.36	Hamartoma	0.85	Patient age	0.31
Leukemia	1.36	Tumor	0.80	Sex	0.12
Bone Dysplasia	1.13	Paternal age	0.77	Rhabdomyoma	0.07
Optic Pathway Glioma	1.09	Epilepsy	0.69	Maternal age	0.05
Cranial Involvement	1.06	Axillary freckling	0.68	Scoliosis	0.02

(Café-au-lait spots: CALS, Cutaneous Neurofibroma: cNF, Plexiform Neurofibroma: pNF, Malignant Peripheral Nerve Sheath Tumor: MPNST)

The performance of the machine learning classifiers on the proposed dataset is presented in Table 3. Precision scores ranged from 0.50 to 0.68, recall values varied between 0.38 and 0.56, F-measure values spanned 0.43 to 0.57, and accuracy scores ranged from 51.43% to 62.86%. The XGBoost method achieved the highest performance, with an accuracy of 62.86%.

Table 3. Performance evaluation of machine learning classifiers on the proposed dataset

Method	Precision	Recall	F-measure	Accuracy (%)
KNN	0.58	0.56	0.57	58.57
ANN	0.61	0.50	0.55	60
SVM	0.59	0.56	0.58	60
DT	0.50	0.38	0.43	51.43
XGBoost	0.68	0.44	0.54	62.86

DISCUSSION

This study evaluated a cohort of NF1 patients from TOBB University, focusing on the differences in clinical symptom frequency between sporadic and familial cases. While the genetic and clinical aspects of NF1 have been extensively studied, there is a notable gap in the literature regarding the specific frequency of symptoms in sporadic versus familial cases.

In this study, machine learning techniques were employed to predict sporadic cases based on clinical symptoms, with the XGBoost algorithm achieving the highest accuracy at 62.86%. While this result demonstrates the potential of machine learning in differentiating between sporadic and familial cases, it also indicates the need for further data collection and feature refinement to improve predictive accuracy. The inclusion of additional clinical features and larger datasets could enhance the model’s performance, allowing for more precise identification of sporadic cases.

In machine learning studies, it is recommended to divide the data set into three parts: learning (training), validation and test data sets. However,our dataset consists of a limited number of samples (121 sporadic and 120 familial cases). Thus, we used for 70% training and 30% testing strategy in this study and we will follow the mentioned strategy in our future studies.

Overall, our findings underscore the complexity of NF1 and the influence of both genetic and environmental factors in its clinical manifestation. Continued research in this area, coupled with advanced analytical techniques, will be essential for developing more accurate diagnostic tools and improving patient outcomes.

In conclusion; this study provides a novel perspective on the phenotypic differences between sporadic and familial NF1 cases, utilizing machine learning to explore these distinctions. While the accuracy of the predictive models was moderate, the findings underscore the complexity of the genotype-phenotype relationship in NF1, suggesting that factors beyond the NF1 gene itself may play a significant role in clinical outcomes. The observed differences in learning disabilities between familial and sporadic cases point to the potential influence of shared genetic or environmental modifiers within families. This research marks a step forward in understanding the variability of NF1 symptoms and highlights the need for more comprehensive data and advanced analytical approaches to improve diagnostic accuracy. Future studies should focus on expanding datasets and refining predictive models to better capture the nuanced interactions between genetic factors and clinical manifestations in NF1 patients. Such efforts will be crucial in advancing personalized medicine and improving patient care in neurogenetic disorders like NF1.

ETHICS COMMITTEE APPROVAL

* The study was approved by the Institutional Ethics Committee of Hacettepe University (Date: 08.11.2016 and Number: 16969557-1117).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

1. Shen MH, Harper PS, Upadhyaya M. Molecular genetics of neurofibromatosis type 1 (NF1). J Med Genet, 1996; 33: 2-17.

2. Rasmussen SA, Friedman JM. NF1 gene and neurofibromatosis 1. Am J Epidemiol, 2000; 151: 33-40.

3. Klose A, Peters H, Hoffmeyer S, Buske A, Luder A, Hess D, et al. Two independent mutations in a family with neurofibromatosis type 1 (NF1). Am J Med Genet, 1999; 83: 6-12.

4. Li Y, O'Connell P, Breidenbach HH, Cawthon R, Stevens J, Xu G, et al. Genomic organization of the neurofibromatosis 1 gene (NF1). Genomics, 1995; 25: 9-18.

5. Upadhyaya M, Majounie E, Thompson P, Han S, Consoli C, Krawczak M, et al. Three different pathological lesions in the NF1 gene originating de novo in a family with neurofibromatosis type 1. Hum Genet, 2003; 112: 12-7.

6. Von Deimling A, Krone W, Menon AG. Neurofibromatosis type 1: pathology, clinical features and molecular genetics. Brain Pathol, 1995; 5: 153-62.

7. Pasmant E, Vidaud M, Vidaud D, Wolkenstein P. Neurofibromatosis type 1: from genotype to phenotype. J Med Genet, 2012; 49: 483-9.

8. Sharafi P, Ayter S. Possible modifier genes in the variation of neurofibromatosis type 1 clinical phenotypes. J Neurogenet, 2018; 32: 65-77.

9. Biotteau M, Dejean S, Lelong S, Iannuzzi S, Faure-Marie N, Castelnau P, et al. Sporadic and familial variants in NF1: an explanation of the wide variability in neurocognitive phenotype? Front Neurol, 2020; 11: 368.

10. Terzi YK, Oguzkan Balci S, Anlar B, Erdogan Bakar E, Ayter S. Learning disability and oligodendrocyte myelin glycoprotein (OMGP) gene in neurofibromatosis type 1. Turk J Pediatr, 2011; 53: 75-8.

11. Sharafi P, Anlar B, Ersoy Evans S, Varan A, Yilmaz OF, Turan M, et al. The effect of parental age on NF1 patients in Turkey. J Community Genet, 2018; 9: 227-32.

12. Terzi YK, Oguzkan Balci S, Anlar B, Varan A, Ersoy Evanse S, Sharafif P, et al. Clinical findings and mutation analysis of NF1 patients in Turkey. Meta Gene, 2018; 15: 80-3.

13. DeBella K, Szudek J, Friedman JM. Use of the national institutes of health criteria for diagnosis of neurofibromatosis 1 in children. Pediatrics, 2000; 105: 608-14.

14. Ferner RE, Huson SM, Thomas N, Moss C, Willshaw H, Evans DG, et al. Guidelines for the diagnosis and management of individuals with neurofibromatosis 1. J Med Genet, 2007; 44: 81-8.

15. Deng Z, Zhu X, Cheng D, Zong M, Zhang S. Efficient kNN classification algorithm for big data. Neurocomputing, 2016; 195: 143-8.

16. Ebiaredoh-Mienye SA, Esenogho E, Swart TG. Integrating enhanced sparse autoencoder-based artificial neural network technique and softmax regression for medical diagnosis. Electronics, 2020; 9: 1963.

17. Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov, 1998; 2:121-67.

18. Breiman L. Random forests. Mach Learn, 2001; 45: 5-32.

19. Rao H, Shi X, Rodrigue AK, Feng J, Xia Y, Elhoseny M, et al. Feature selection based on artificial bee colony and gradient boosting decision tree. Appl Soft Comput, 2019; 74: 634-42.

20. Terzi YK, Sirin B, Serdaroglu E, Anlar B, Aysun S, Hosgor G, et al. Absence of exon 17 c.2970-2872 del AAT mutation in Turkish NF1 patients with mild phenotype. Childs Nerv Syst, 2011; 27: 2113-6.