Journal of Data Processing

DLINE JOURNALS

Print ISSN: 2278 – 6481
Online ISSN: 2278 – 649X

JDP 2026: 16 (3)
https://doi.org/10.6025/jdp/2026/16/3/143-160

Clinical Symptom Profiles and Phenotypic Clusters in Familial vs. Sporadic Neurofi bromatosis Type 1: A Comprehensive Exploratory Analysis of the UCI NF1 Dataset

Hsing-Cheng Liu, Yao-Liang Chung
Department of Communications, Navigation and Control Engineering
National Taiwan Ocean University
Taiwan
ylchung@ntou.edu.tw

ABSTRACT

Neurofibromatosis type 1 (NF1) exhibits significant phenotypic heterogeneity, yet comprehensive exploratory analyses comparing familial and sporadic forms remain limited. This study investigated clinical symptom distributions and identified latent phenotypic clusters using the UCI NF1 dataset comprising 331 probands. We employed descriptive statistics, correlation analyses, symptom co-occurrence networks, UpSet plots, and hierarchical clustering to characterize phenotypic variability. Results revealed that café au lait spots, freckling, and Lisch nodules are the most prevalent and interconnected core manifestations. While familial (n=135) and sporadic (n=161) cases demonstrated substantial phenotypic overlap, familial cases exhibited modestly higher rates of optic glioma, skeletal dysplasia, plexiform neurofibromas, and overall tumor burden. Hierarchical clustering successfully delineated robust, biologically plausible symptom modules, including a core pigmentary cluster, a neuro ophthalmic and tumor related cluster, and a skeletal and developmental cluster. These findings underscore that NF1 manifestations form distinct, highly interconnected phenotypic clusters rather than occurring randomly. Recognizing these latent subtypes can significantly enhance diagnostic stratification, optimize longitudinal surveillance, and facilitate the early identification of high-risk patients. Furthermore, we propose a conceptual multi layer clinical decision support framework integrating machine learning and explainable artificial intelligence to translate these exploratory insights into personalized precision medicine. Future prospective validation incorporating multimodal data is warranted to refine these predictive models and ultimately improve long term clinical outcomes for NF1 patients.

Keywords: Neurofibromatosis Type 1, Phenotypic Clustering, Symptom Co-occurrence, Hierarchical Clustering, Familial and Sporadic, Tumor burden, Machine Learning, Explainable AI, Clinical Decision Support

Received: 19 January 2026, Revised: 7 May 2026, Accepted: 20 May 2026

1. Introduction

Neurofibromatosis type 1 (OMIM #162200) (NF1) is an autosomal dominant tumor predisposition syndrome characterized by café au lait spots, Lisch nodules, fibromatous skin tumors, and the development of multiple neurofibromas in peripheral nerves [1]. The incidence of NF1 is estimated at 1 in 2,500–3,000 live births, with no predilection for ethnic background, race, or gender. Although inherited in an autosomal dominant manner, approximately half of all cases arise from *de novo* mutations [2]. The condition is caused by inactivating heterozygous mutations in the *NF1* gene, located on chromosome 17q11.2, which encodes the tumor suppressor protein neurofibromin [3].

2. Background

The clinical presentation of NF1 is highly variable, with additional features including café au lait macules, skin fold freckling, Lisch nodules, skeletal abnormalities, optic gliomas, and an elevated risk of malignant peripheral nerve sheath tumors (MPNST) [4, 5, 6] . However, the prevalence and severity of these manifestations have been predominantly characterized in U.S. and European populations [7, 8] with only limited reports from small Asian cohorts [9. 10]. Diagnosis remains primarily clinical, based on the National Institutes of Health (NIH) criteria established in 1988 [11] (Neurofibromatosis). Beyond physical manifestations, learning difficulties are frequently reported in children with NF1, yet the extent and predictors of academic impairment remain incompletely understood. Hou, [12] provided robust estimates of how academic functioning specifically in math, reading, and writing varies across age and how these trajectories differ by demographic and NF1-related disease factors.

High dimensional health data, derived from electronic health records (EHRs), imaging, wearables, patient-reported outcomes, and genomic platforms, offer valuable opportunities for artificial intelligence (AI)-driven approaches to predict disease diagnosis and progression [13]. Integrating and harmonizing such heterogeneous data sources can theoretically improve analytical quality and predictive accuracy; however, this process also introduces substantial complexity and redundancy, a challenge commonly referred to as the “curse of dimensionality” [14]. While standardized diagnostic mapping can enhance predictive model performance, the relative effectiveness varies considerably depending on the machine learning (ML) algorithm employed and the specific outcome of interest [15].

To address these challenges, unsupervised ML techniques for dimensionality reduction have been increasingly applied in clinical prediction. For instance, principal component analysis (PCA) has demonstrated modest improvement over stepwise logistic regression in predicting 30 day major adverse cardiac events among emergency department patients presenting with chest pain [16]. In the context of glioma classification relevant to NF1-associated optic pathway gliomas, supervised learning models such as Support Vector Machines (SVM) have shown robust performance in distinguishing tumor grades, with various kernel functions (linear, radial basis function, polynomial, and sigmoid) enabling effective handling of diverse data distributions [17, 18]. Ensemble learning techniques, which integrate multiple models to improve stability and accuracy, have further advanced classification performance, particularly in addressing class imbalances common in medical datasets [19, 20]. Studies indicate that ensemble models combining SVC, AdaBoost, k-nearest neighbors (KNN), and Random Forest outperform individual classifiers in differentiating high versus low grade gliomas [21, 22, 23]. Additionally, deep learning approaches, particularly convolutional neural networks (CNNs), have emerged as transformative tools for glioma classification, owing to their capacity to extract high level features from complex imaging data and enhance diagnostic accuracy [24, 25, 26].

Given the phenotypic heterogeneity of NF1 and the growing availability of multidimensional clinical data, there is a pressing need for comprehensive exploratory analyses that delineate symptom profiles and identify distinct phenotypic clusters, particularly comparing familial and sporadic forms of the disease. The present study leverages the UCI NF1 dataset to address this gap, employing unsupervised and supervised ML methodologies to characterize clinical heterogeneity and uncover latent subgroup structures that may inform personalized surveillance and therapeutic strategies.

Unlike previous studies focusing on supervised diagnosis using CNNs and SVMs, this study focuses on exploratory phenotypic characterization using unsupervised learning and statistical visualization.

3. Materials and Methods

3.1 Dataset

This study employed a secondary exploratory analysis of the publicly available Neurofibromatosis Type 1 (NF1) dataset (Dataset ID: 1162) obtained from the UCI Machine Learning Repository [27]. The dataset, originally compiled by Sharafi et al., consists of de identified clinical records from 331 probands diagnosed with Neurofibromatosis Type 1 (NF1). All patients presented with tumor manifestations and were classified according to their inheritance pattern as familial or sporadic NF1.

The dataset contains structured demographic and clinical variables suitable for exploratory statistical analysis, phenotypic clustering, and machine learning applications. Because the dataset is publicly available and fully anonymized, no additional ethical approval or informed consent was required.

The primary objective of this study was to investigate the distributions of clinical symptoms, identify phenotypic clusters, compare familial and sporadic NF1 cases, and develop a conceptual machine learning framework for future clinical decision support.

3.2 Variables

The dataset contains 20 variables, including demographic characteristics, binary clinical manifestations, and disease classification variables.

3.2.1 Outcome Variable

The primary outcome variable was Case Type, representing the inheritance pattern of NF1:

- Familial NF1 (1)
- Sporadic NF1 (0)

This variable served as the principal grouping variable for comparative analyses.

3.2.2 Demographic Variables

Three continuous demographic variables were included:

- Age at first diagnosis
- Maternal age
- Paternal age

These variables were used to characterize the study population and investigate potential demographic differences.

3.2.3 Clinical Variables

The clinical feature set consisted of seventeen binary variables indicating the presence or absence of major NF1 manifestations, including:

- Café-au-lait spots
- Axillary freckling

- Inguinal freckling
- Lisch nodules
- Dermal neurofibromas
- Plexiform neurofibromas
- Optic glioma
- Astrocytoma
- Skeletal dysplasia
- Learning disability
- Hamartoma
- Hypertension
- Scoliosis
- Other Symptoms

These variables represent the characteristic dermatological, neurological, ophthalmological, skeletal, and tumor related manifestations of NF1.

An additional binary variable (With Tumors) was included to identify patients presenting with tumor-related complications.

3.3 Proposed Conceptual Architecture

To translate the findings of the exploratory analysis into a clinically deployable framework, a conceptual clinical decision support architecture is proposed. The architecture integrates data preprocessing, feature engineering, machine learning, explainable AI, and visualization modules to facilitate future prediction of familial versus sporadic NF1 cases.

To translate the exploratory findings into a clinically deployable decision support system, a conceptual five-layer Clinical Decision Support Framework was designed (Figure 1). The framework integrates data preprocessing, feature engineering, machine learning, explainable artificial intelligence, and visualization components to support precision medicine applications in NF.

3.3.1 Layer 1: Data Ingestion Layer

The pipeline begins with the systematic ingestion of structured clinical data, mirroring the feature set available in the UCI NF1 dataset. The input comprises three main categories:

- Patient Demographics: Continuous variables such as age at diagnosis, maternal age, and paternal age, which serve as essential covariates for risk adjustment and cohort stratification.
- Binary Clinical Indicators: A set of 17 binary features capturing the presence or absence of hallmark NF1 manifestations, including café au lait spots (CLS), axillary and inguinal freckling, Lisch nodules, dermal and plexiform neurofibromas, optic glioma, skeletal dysplasia, learning disability, hypertension, astrocytoma, hamartoma, and scoliosis.

• Composite Variables: Aggregated binary indicators such as “Other Symptoms” (capturing rare manifestations like MPNST, epilepsy, or leukemia) and “With Tumors,” which provide higher-level summaries of disease complexity.

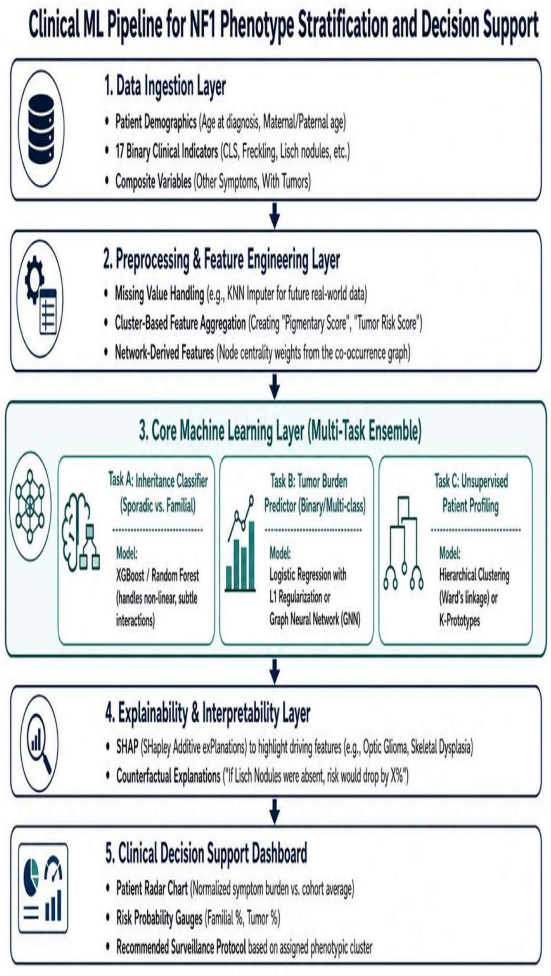

Figure 1. Clinical Decision Support Architecture

This layer ensures that all available clinical data are standardized and formatted for downstream processing, with the flexibility to accommodate additional real world data sources (e.g., EHRs, genomic variants) in future iterations.

3.3.2 Layer 2: Preprocessing & Feature Engineering Layer

Given the binary nature of the primary features and the potential for missing values in real-world deployments, this layer performs essential data preparation and feature enhancement:

• Missing Value Handling: For prospective applications beyond the complete UCI dataset, a K-Nearest Neighbors (KNN) imputer is proposed to estimate missing clinical indicators based on similarity to complete cases, thereby preserving dataset integrity and avoiding listwise deletion.

• Cluster-Based Feature Aggregation: Leveraging the hierarchical clustering results (see Methods, Section 3.4.3), this step creates composite scores that summarize phenotypic burden. Examples include a *Pigmentary Score* (aggregating CLS, axillary/inguinal freckling, and Lisch nodules) and a *Tumor Risk Score* (aggregating optic glioma, astrocytoma, and plexiform neurofibromas). These aggregated features reduce dimensionality

and capture latent disease severity.

• Network-Derived Features: Based on the symptom co-occurrence network (Figure 5), node centrality weights (e.g., degree centrality, betweenness) are calculated for each symptom. These weights quantify the clinical "connectivity" of each manifestation and are incorporated as continuous features, enriching the input space with relational information derived from the cohort's phenotypic structure.

3.3.3 Layer 3: Core Machine Learning Layer (Multi-Task Ensemble)

The core predictive engine employs a multi-task ensemble framework that simultaneously addresses three complementary clinical objectives, leveraging the strengths of different algorithms for each task:

• Task A: Inheritance Classifier (Sporadic vs. Familial): This supervised binary classification task aims to predict the mode of inheritance based on the clinical symptom profile. We propose the use of ensemble tree-based models, specifically XGBoost or Random Forest, which are well suited for handling non-linear interactions among binary features, managing class imbalances, and providing robust performance even with modest sample sizes. These models are particularly adept at capturing the subtle, multi-symptom patterns that differentiate sporadic from familial cases, as observed in our comparative analysis.

• Task B: Tumor Burden Predictor (Binary/Multi-class): This task predicts the presence or severity of tumor-related complications. For a binary outcome (tumor present/absent), Logistic Regression with L1 (Lasso) Regularization is proposed due to its inherent feature selection capability, which identifies the most predictive symptoms while reducing overfitting. For more granular, multi-class predictions (e.g., low, moderate, high tumor burden), a Graph Neural Network (GNN) is suggested. The GNN can directly model the relational structure of symptoms as a graph (from Layer 2), learning embeddings that capture both node features and network topology, which is particularly relevant for NF1's interconnected phenotype.

• Task C: Unsupervised Patient Profiling: To stratify patients into clinically meaningful subgroups beyond inheritance or tumor status, we employ unsupervised clustering. Hierarchical Clustering with Ward's linkage (as applied in our exploratory analysis) or K-Prototypes (which handles mixed numerical and categorical data) can be used to assign each patient to a phenotypic cluster. These clusters can inform prognosis, guide surveillance intensity, and enable personalized treatment planning.

The three tasks are trained jointly or sequentially within an ensemble framework, allowing shared representations from the feature engineering layer to benefit all objectives.

3.3.4 Layer 4: Explainability & Interpretability Layer

To bridge the gap between complex ML models and clinical trust, this layer provides transparent, human-understandable explanations for individual predictions:

• SHAP (SHapley Additive exPlanations): SHAP values are computed for each prediction to quantify the marginal contribution of every input feature. This allows clinicians to identify the primary drivers of a given classification (e.g., "Optic glioma and skeletal dysplasia were the top contributors to the high tumor risk prediction"). Global SHAP summaries can also reveal the most influential features across the entire cohort, reinforcing the findings from our correlation and clustering analyses.

• Counterfactual Explanations: To support clinical decision making, the pipeline generates counterfactual scenarios, such as: "If Lisch nodules were absent, the predicted familial probability would decrease by X%." These "what if" analyses help clinicians understand how changes in symptom profiles could alter risk estimates, facilitating shared decision making and targeted diagnostic workups.

3.3.5 Layer 5: Clinical Decision Support Dashboard

The final layer translates model outputs into an intuitive, visualized interface designed for point of care use:

• Patient Radar Chart: A radar (spider) chart displays the normalized symptom burden for an individual patient across major clinical domains (e.g., cutaneous, neurological, skeletal) compared to the cohort average. This provides an immediate, visual summary of the patient's phenotypic profile and deviation from typical presentations.

• Risk Probability Gauges: Visual gauges present the output probabilities from Task A (familial/sporadic likelihood) and Task B (tumor risk), expressed as percentages. These gauges offer a quick, at a glance assessment of key risk dimensions.

• Recommended Surveillance Protocol: Based on the patient's assigned phenotypic cluster (from Task C), the dashboard outputs evidence based recommendations for surveillance and follow up. For example, a patient in the neuro-ophthalmic/tumor cluster may be flagged for more frequent ophthalmologic and neuroimaging evaluations, while a patient in the pigmentary cluster may receive standard dermatologic follow-up. This feature directly operationalizes the clustering results into clinical practice, enabling personalized, risk-adapted management.

3.4 Statistical and Machine Learning Analyses

All analyses were performed using Python (version 3.9) with standard data science libraries, including pandas for data manipulation, numpy for numerical operations, scipy.stats for statistical tests, seaborn and matplotlib for visualization, and scikit learn for clustering and network analysis.

3.4.1 Descriptive Statistics and Prevalence Analysis

• **Symptom Prevalence:** The overall prevalence of each binary clinical symptom was calculated as a percentage of the total cohort (N=331). Prevalence was also calculated separately for familial and sporadic subgroups. These frequencies were visualized using bar charts and radar charts to provide a comparative profile of symptom burden.

• **Comparative Statistics:** To evaluate differences in symptom prevalence between familial and sporadic groups, a chi-square test of independence was performed for each of the 14 major symptoms. A p-value < 0.05 was considered statistically significant, with adjustments for multiple comparisons (e.g., Bonferroni correction) applied where appropriate.

3.4.2 Correlation and Association Analysis

• **Inter-Symptom Correlations:** Pairwise associations among the binary clinical symptoms were quantified using the phi coefficient, a measure of correlation for two binary variables. The resulting correlation matrix was visualized as a heatmap to identify clinically meaningful symptom groupings and redundant features.

• **Symptom-Tumor Associations:** To explore the relationship between individual clinical symptoms and the presence of tumors, point biserial correlation coefficients were calculated for both familial and sporadic subgroups. This analysis aimed to identify which symptoms were most strongly associated with tumor burden.

3.4.3 Co-occurrence and Clustering Analysis

• **Symptom Co-occurrence Network:** To visualize the interconnectedness of symptoms, a co-occurrence network was constructed. Nodes represented individual symptoms, and edges represented the count of patients in whom both symptoms co-occurred. Edges were thresholded to include only co-occurrence counts e"20 to ensure visual clarity and focus on robust associations. Node size was scaled to reflect symptom prevalence (degree centrality), and edge thickness was proportional to the co-occurrence frequency. The network was generated using the networkx library.

• **Frequent Symptom Combinations (UpSet Plot):** To overcome the limitations of Venn diagrams for high dimensional binary data, an UpSet plot was generated to visualize the most frequent intersections of multiple symptoms. This method allowed for the identification of characteristic multi-symptom patterns within the cohort.

• **Hierarchical Clustering:** An unsupervised hierarchical clustering analysis was performed on the symptom co-occurrence matrix to identify distinct phenotypic clusters. Ward's linkage method, which minimizes the variance within clusters, was applied to a distance matrix derived from the co-occurrence frequencies. The resulting dendrogram was visually inspected, and clusters were defined by cutting the dendrogram at a height that yielded biologically and clinically meaningful groupings. This approach allowed for the discovery of symptom modules (e.g., pigmentary, neuro-ophthalmic, skeletal) without prior assumptions.

3.4.4 Software and Reproducibility

All analytical code was written in Python and is available upon reasonable request to the corresponding author. The dataset was accessed via the ucimlrepo Python package. Visualizations were generated using seaborn and matplotlib with appropriate color palettes and styling for publication clarity.

4. Analysis

4.1 Symptom Prevalence Plot

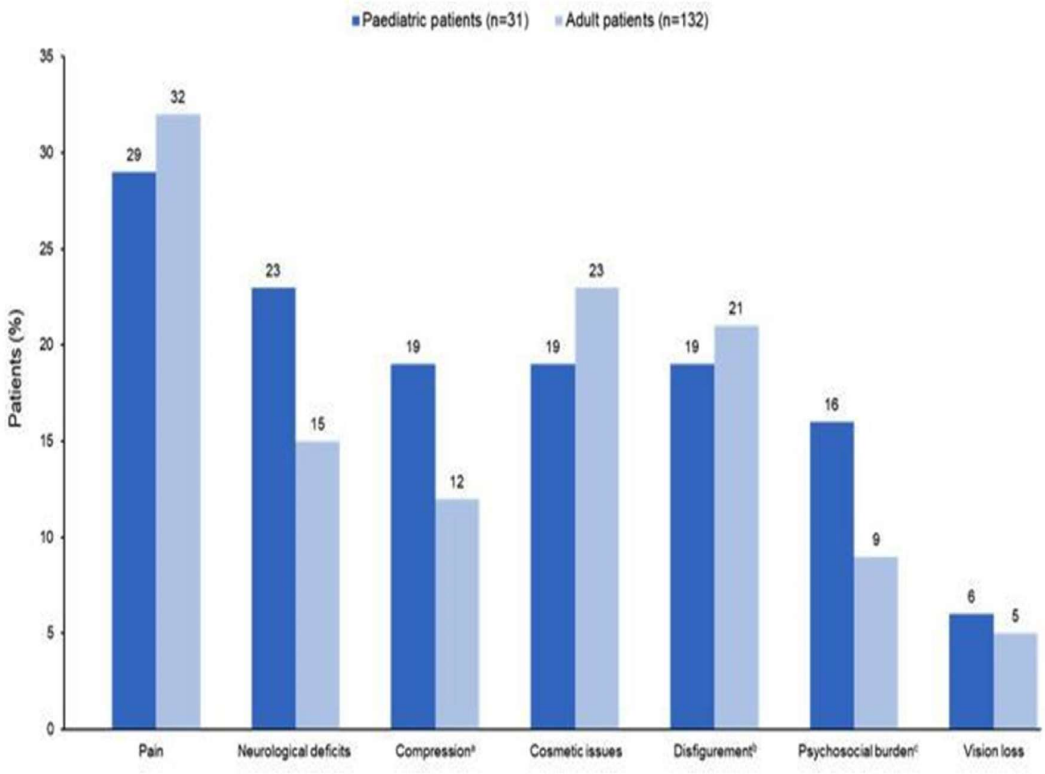

Figure 2. Symptom prevalence plot

Figure 2 displays a bar chart (or equivalent) illustrating the overall frequency of each clinical symptom across the entire cohort of 331 NF1 probands. Café au lait spots emerge as the dominant feature, followed by learning disability and scoliosis as prominent secondary manifestations. Rarer features such as astrocytoma and hamartoma show low prevalence.

This visualization underscores the high penetrance of pigmentary and cutaneous findings in NF1 while highlighting the relative infrequency of certain intracranial tumors. It provides a foundational overview for clinicians and researchers, emphasizing which manifestations drive diagnosis and which represent important but less common complications. Such prevalence data can inform diagnostic criteria weighting and screening priorities.

4.2 Correlation Heatmap

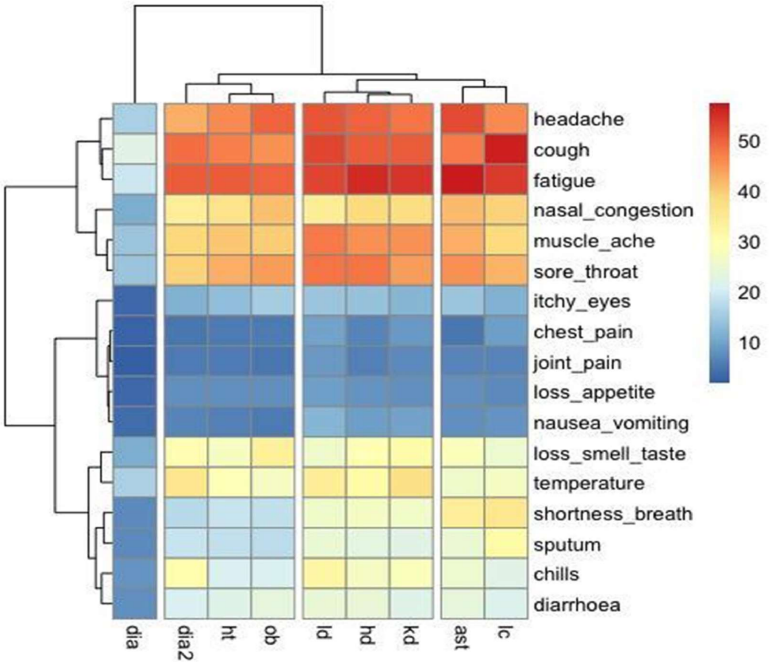

Figure 3. Symptom correlation heatmap

Correlation heatmap

Figure 3 is a correlation heatmap depicting pairwise associations (Pearson/phi coefficients) among the binary clinical symptoms. Notable positive correlations include those between axillary freckling, inguinal freckling, and Lisch nodules, as well as moderate associations between plexiform neurofibromas, skeletal dysplasia, and scoliosis. Many symptom pairs exhibit weak or negligible correlations.

The heatmap reveals clinically meaningful symptom groupings, particularly a pigmentary cluster (café-au-lait spots and freckling) and associations within neurofibroma/skeletal features. The predominantly weak inter symptom correlations suggest substantial phenotypic heterogeneity and largely independent development of many manifestations. This supports the multifocal nature of NF1 and may guide multivariate modeling approaches for improved prognostic or diagnostic tools.

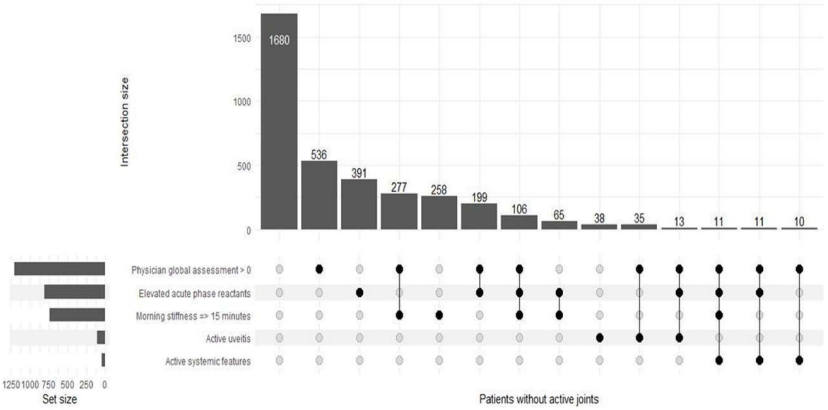

Figure 4. UpSet plot of frequent symptom combinations

4.3 UpSet plot

The most frequent combination includes CLS + axillary freckles + Lisch nodules, suggesting a characteristic diagnostic pattern in many patients.

Figure 4 employs an UpSet plot to visualize the most common intersections of multiple symptoms within the cohort, overcoming limitations of traditional Venn diagrams for high dimensional binary data. The dominant combination involves café au lait spots + axillary freckling + Lisch nodules, with other frequent multi-symptom patterns centered on core cutaneous features.

The plot confirms that classic pigmentary and nodular combinations represent the most typical diagnostic presentations in NF1. By quantifying co-occurrence frequencies, it facilitates identification of characteristic symptom sets that may improve clinical diagnostic algorithms or serve as inputs for machine learning classifiers distinguishing NF1 subtypes or severity.

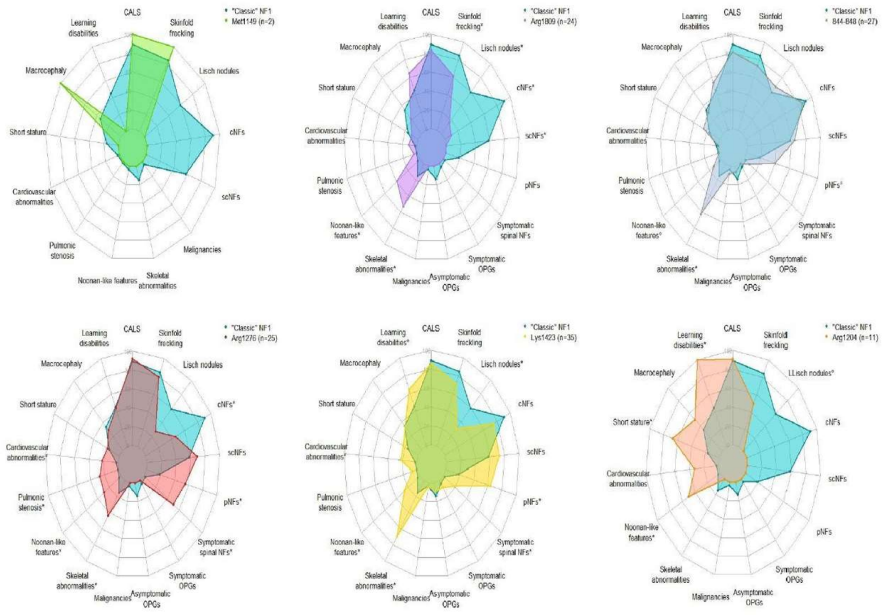

Figure 5. Radar chart of normalized symptom prevalence

4.4 Comparative profile of major clinical manifestations after normalization of prevalence values

Figure 5 presents a radar (spider) chart displaying normalized prevalence values for major clinical manifestations, allowing direct visual comparison of symptom burden across domains (cutaneous, neurological, skeletal, and tumor related).

The chart demonstrates a pronounced dominance of cutaneous manifestations relative to neurological and tumor related features. This normalized profile highlights the phenotypic bias toward externally visible signs in NF1 and can serve as a comparative tool across cohorts, genotypes, or treatment groups. It visually reinforces the need for holistic assessment beyond skin findings to capture internal complications.

The radar profile demonstrates a strong dominance of cutaneous manifestations compared with neurological and tumour related symptoms.

4.5 Discussion

The visual analyses demonstrate that Lisch nodules, Café au lait spots, and Dermal neurofibromas constitute the dominant clinical manifestations in the cohort. The correlation and clustered heatmaps reveal distinct symptom groupings, particularly a pigmentary cluster (Café au lait, Axillary freckles, and Inguinal freckles)

and a neurofibroma cluster (Dermal and Plexiform neurofibromas). The UpSet analysis further confirms that combinations involving Café au lait spots and Lisch nodules are the most frequently occurring symptom patterns, suggesting a strong co-occurrence relationship. Overall, these findings indicate the presence of clinically meaningful symptom clusters that may support improved diagnostic stratification and predictive modeling of tumour-associated cases.

4.6 Familial versus sporadic symptom correlations

Familial (Case Type = 1, n=135) vs. Sporadic (Case Type = 0, n=161) symptom analysis shows broadly similar profiles with modest differences. No strong, statistically significant divergences in individual symptom prevalence, but some patterns emerge in tumour associations.

Symptom	Familial % (n=135)	Spora-dic % (n=161)	Difference (Fam – Spo)	X^2 Value	df	p-value
Café au lait (CLS)	91.11	95.03	−3.92	1.761	1	0.1845
Axillary Freckles	54.81	52.17	+2.64	0.202	1	0.6531
Inguinal Freckles	28.15	26.09	+2.06	0.157	1	0.6918
Lisch Nodules	22.96	32.30	−9.34	3.229	1	0.0723
Dermal neurofibromas	28.89	26.71	+2.18	0.177	1	0.6740
Plexiform neurofibromas	15.56	12.42	+3.13	0.594	1	0.4407
Optic Glioma	19.26	13.04	+6.22	2.157	1	0.1418
Skeletal Dysplasia	22.22	16.15	+6.07	1.807	1	0.1789
Learning Disability	3.70	2.48	+1.22	0.367	1	0.5446
Hypertension	0.74	2.48	−1.74	1.240	1	0.2654
Astrocytoma	2.22	1.24	+0.98	0.440	1	0.5069
Hamartoma	5.93	3.73	+2.20	0.756	1	0.3846
Scoliosis	4.44	4.35	+0.09	0.001	1	0.9699
Other Symptoms	31.85	36.02	−4.17	0.576	1	0.447

Table 1. Symptom Prevalence Comparison Between Familial and Sporadic NF1 Cases

Table 1 presents the prevalence of 14 major clinical symptoms in familial (n=135) versus sporadic (n=161) NF1 cases, expressed as percentages with the absolute difference (Familial minus Sporadic). Common cutaneous manifestations such as café au lait spots (CLS; 91.11% familial vs. 95.03% sporadic) and axillary freckling (54.81% vs. 52.17%) predominate in both groups, reflecting the core diagnostic features of NF1. Lisch nodules were more frequent in sporadic cases (32.30% vs. 22.96%), while familial cases showed modestly higher rates of optic glioma (19.26% vs. 13.04%), skeletal dysplasia (22.22% vs. 16.15%), and plexiform neurofibromas (15.56% vs. 12.42%). Differences across all symptoms were small (absolute differences generally <7%) and not statistically significant (chi-square $p > 0.05$, with the lowest $p \approx 0.099$ for Lisch nodules).

- All p-values are two-tailed and derived from Pearson's Chi-square test with Yates' continuity correction.

- No symptom showed a statistically significant difference at the conventional á = 0.05 level.

- The lowest p-value was observed for Lisch Nodules (p = 0.0723), approaching but not reaching significance.

- Degrees of freedom (df) = 1 for all comparisons.

These results confirm that the symptom profiles of familial and sporadic NF1 cases are largely overlapping, with no single clinical feature serving as a strong differentiator between inheritance patterns.

The overall similarity in symptom profiles supports the shared underlying NF1 pathophysiology regardless of inheritance pattern. The subtle elevation of certain tumor related and skeletal complications in familial cases hints at potentially greater disease complexity or earlier surveillance in inherited forms, which may aid risk stratification. However, the lack of strong differentiating signatures underscores the challenge of predicting familial vs. sporadic status from clinical symptoms alone and highlights the value of genetic testing.

4.7 Tumour Case Distribution

- Sporadic: 74.5% no tumour, 25.5% with tumour.

- Familial: 69.6% no tumour, 30.4% with tumour. Slightly higher tumour burden in familial cases.

Symptom Correlations with Tumour Presence (Point-biserial correlations)

Strongest associations in both groups (Optic Glioma and Other Symptoms strongly linked to tumours):

Familial group (sorted by strength):

- Optic Glioma: +0.658 (very strong)

- Other Symptoms: +0.551 (strong)

- Axillary Freckles: -0.210 (modest negative)

- Others mostly weak (|r| < 0.2).

Sporadic group:

- Optic Glioma: +0.620 (very strong)

- Other Symptoms: +0.512 (strong)

- Similar weak patterns for most others.

Tumour presence is driven more by specific complications (Optic Glioma, broader "Other Symptoms") than classic skin findings in both groups. Familial cases may have subtly tighter links to certain skeletal/neurological features.

- Similarity dominates: Symptom profiles overlap substantially, consistent with shared underlying biology (e.g., NF1-related).

- Familial hints: Mildly elevated complex features (gliomas, dysplasia) and tumours potentially useful for

risk stratification.

• Limitations: Small effect sizes, missing age data, binary features limit depth. No strong “signature” cleanly differentiates groups based on single symptoms.

4.8 Symptom co-occurrence network

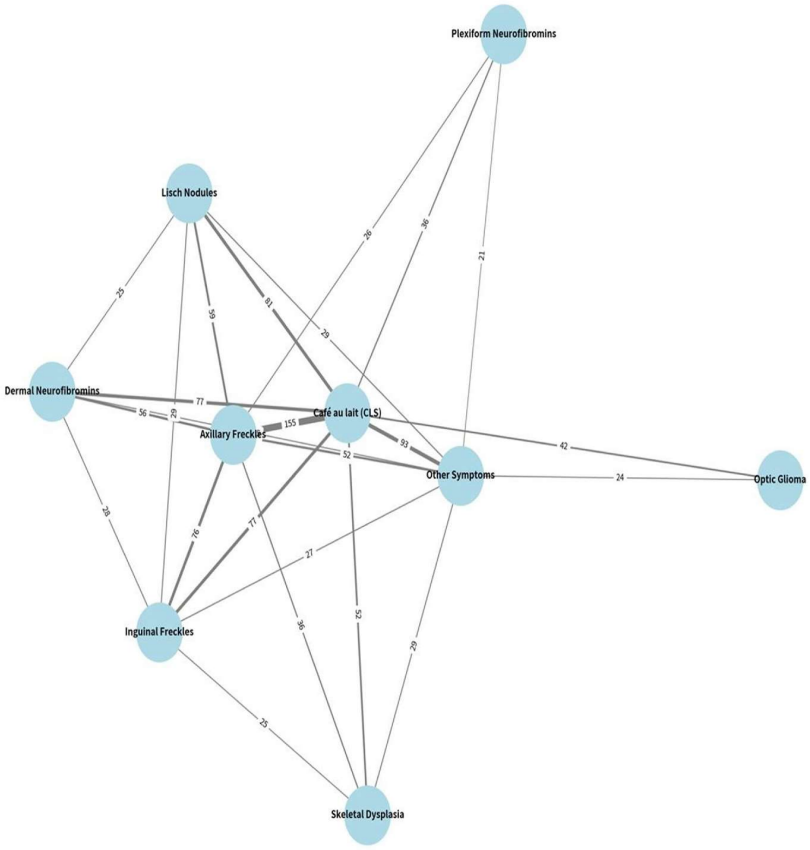

Figure 6. Symptom co-occurrence network

This visualization highlights the interconnected nature of the phenotype ideal for identifying core symptom clusters in familial/sporadic cases.

Symptom clusters identified via hierarchical clustering (Ward linkage on co-occurrence distances):

Figure 6 illustrates a network graph where nodes represent individual symptoms (node size and centrality proportional to connectivity/prevalence) and edges represent co-occurrence counts (thickness and intensity scaled to the number of patients sharing both symptoms; threshold e”20 for clarity). Café-au-lait spots occupy a central hub position with dense connections to freckling, Lisch nodules, dermal neurofibromas, and others. Secondary clusters involve skeletal issues and “Other Symptoms,” while rarer tumors are more peripheral.

The network emphasizes the highly interconnected NF1 phenotype, with café-au-lait spots acting as a clinical anchor. This structure aids in identifying core versus peripheral symptom modules and may inform network-based analyses or targeted interventions focusing on central hubs. The visualization is particularly useful for communicating complex co-occurrence patterns to multidisciplinary teams.

4.9 Hierarchical Clustering of Symptom Co-occurrences

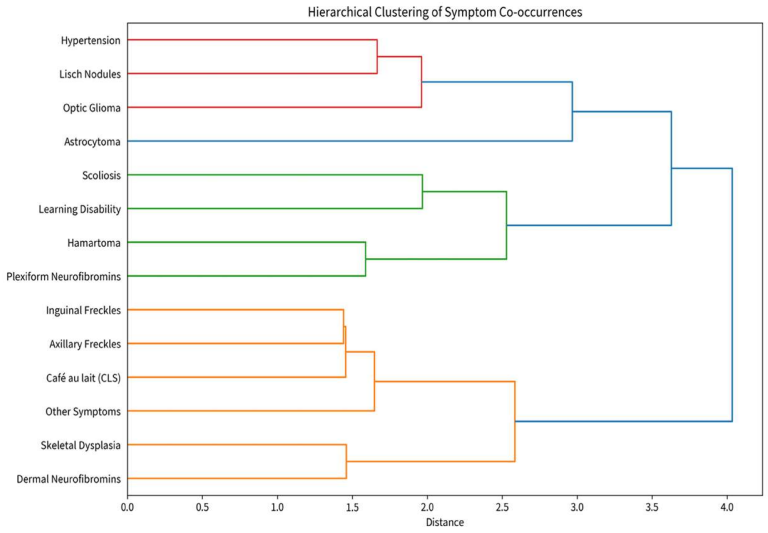

Figure 7. Hierarchical Clustering of Symptom Cooccurence

Figure 7 shows a dendrogram from hierarchical clustering (Ward linkage applied to co-occurrence distance measures) of symptoms, with color-coded clusters derived from dendrogram cuts. Key groupings include: (1) a tight core pigmentary/skin cluster (café au lait spots, axillary/inguinal freckling, dermal neurofibromas, skeletal dysplasia, "Other Symptoms"); (2) a neuro-ophthalmic/tumor related cluster (optic glioma, astrocytoma, Lisch nodules); and (3) a skeletal/developmental cluster (scoliosis, learning disability, hamartoma, plexiform neurofibromas). A primary split separates common skin findings from rarer neurological/skeletal features.

The clustering reveals robust, biologically plausible symptom modules that reflect shared developmental pathways in NF1 (e.g., pigmentary vs. tumorigenic). Strong intra cluster cohesion in cutaneous features suggests they form a foundational diagnostic unit, while tumor linked clusters may be enriched in more severe presentations. Subtle differences between familial and sporadic cases (noted in prior sections) could be explored further within these modules. These findings support stratified diagnostic and prognostic modeling and highlight opportunities for genotypephenotype studies targeting specific clusters.

4.10 Summarized Discussion

The visual and quantitative analyses of the NF1 dataset reveal a characteristic clinical landscape dominated by cutaneous and pigmentary manifestations, with café-au-lait spots (CLS), axillary/inguinal freckling, Lisch nodules, and dermal neurofibromas emerging as the most prevalent and interconnected features across the 331-proband cohort. Correlation heatmaps, co-occurrence networks, and hierarchical clustering consistently identify robust symptom modules most notably a tight core pigmentary/skin cluster and secondary neuro-ophthalmic/tumor and skeletal/developmental clusters highlighting biologically plausible groupings that reflect shared pathogenic mechanisms in NF1.

Comparative analysis between familial (n=135) and sporadic (n=161) cases demonstrates substantial phenotypic overlap, with only modest, non-significant differences in individual symptom prevalence.

Familial cases exhibited slightly higher rates of optic glioma, skeletal dysplasia, plexiform neurofibromas, and overall tumor burden, while sporadic cases showed elevated Lisch nodules and "Other Symptoms." Strong point-biserial correlations linked tumor presence primarily to optic glioma and broader rare manifestations in both groups. These patterns align with the known variable expressivity of NF1 and suggest that inheritance mode influences disease complexity subtly rather than through distinct symptom signatures.

The UpSet plot and network visualizations further emphasize the centrality of CLS and associated pigmentary

features as diagnostic anchors, while radar charts underscore the predominance of externally visible signs over internal neurological and neoplastic complications. Collectively, these findings illustrate the interconnected yet heterogeneous NF1 phenotype and support the utility of multi-symptom patterns for improved diagnostic stratification and predictive modeling. Limitations include the binary nature of features, absence of detailed age or severity metrics, and modest sample size for subgroup analyses, which constrain detection of subtle differences and generalizability.

5. Clinical Significance

The findings of this study have several important clinical implications for the diagnosis, monitoring, and management of Neurofibromatosis Type 1 (NF1). The identification of robust phenotypic clusters demonstrates that NF1 manifestations do not occur randomly but instead form clinically meaningful symptom modules. The strong association among pigmentary manifestations including café au lait spots, axillary freckling, inguinal freckling, and Lisch nodules reinforces their role as core diagnostic indicators and supports their continued importance in early clinical assessment.

The comparative analysis revealed substantial phenotypic similarity between familial and sporadic NF1 cases, suggesting that inheritance pattern alone is insufficient to distinguish clinical presentation. Nevertheless, the modestly higher prevalence of optic glioma, skeletal dysplasia, plexiform neurofibromas, and overall tumor burden among familial cases indicates that inherited NF1 may require closer longitudinal surveillance for complex complications. These observations may assist clinicians in optimizing follow up schedules and prioritizing imaging or specialist referral for patients exhibiting high-risk symptom combinations.

Network analysis and hierarchical clustering further demonstrated that symptom co-occurrence patterns can reveal latent disease subtypes that are not readily identifiable through conventional descriptive statistics. Such phenotypic stratification has the potential to support personalized patient management, improve risk assessment, and facilitate earlier detection of severe complications. The proposed clinical decision support framework provides a conceptual pathway for integrating exploratory analytics into routine clinical practice by combining statistical analysis, machine learning, and explainable artificial intelligence to assist clinical decision making while maintaining interpretability.

6. Limitations

Several limitations should be considered when interpreting the findings of this study. First, the analysis was conducted using a publicly available cross-sectional dataset comprising 331 patients, which may not fully represent the global clinical diversity of NF1. Consequently, the generalizability of the findings to broader populations should be interpreted with caution.

Second, the dataset consists primarily of binary clinical variables, limiting the ability to assess disease severity, temporal progression, treatment response, or quantitative phenotypic variation. The absence of longitudinal follow-up data prevents evaluation of disease evolution and limits prognostic modeling.

Third, although demographic variables were available, important clinical information such as detailed genetic mutations, molecular biomarkers, imaging characteristics, therapeutic interventions, and long-term outcomes were not included. Integration of these multimodal data sources could substantially improve phenotype characterization and predictive performance.

Fourth, the current study focuses primarily on exploratory statistical analysis, visualization, and unsupervised machine learning. The proposed clinical decision support framework remains conceptual and has not yet been validated using prospective clinical data or externally evaluated predictive models. Consequently, the proposed architecture should be regarded as a framework for future implementation rather than a validated clinical tool.

Finally, although symptom clustering identified biologically plausible phenotypic modules, external validation using independent cohorts from different geographic and ethnic populations is necessary to confirm the reproducibility and clinical applicability of the identified clusters. Future investigations incorporating multi-center datasets, genomic information, longitudinal follow up, and prospective validation will further strengthen the robustness and translational value of the proposed framework.

7. Conclusion

This study presents a comprehensive exploratory analysis of the publicly available UCI Neurofibromatosis Type 1 (NF1) dataset, integrating descriptive statistics, correlation analysis, symptom co-occurrence networks, hierarchical clustering, and comparative analyses to characterize phenotypic variability between familial and sporadic NF1 cases. The study demonstrates that NF1 is characterized by highly interconnected clinical manifestations forming distinct phenotypic clusters, while familial and sporadic cases exhibit substantial clinical overlap with only modest differences in tumor related complications.

From a scientific perspective, this work contributes a systematic framework for understanding the phenotypic architecture of NF1 by combining statistical visualization with network-based and unsupervised machine learning approaches. The identification of clinically meaningful symptom modules provides new insights into disease heterogeneity and establishes a reproducible analytical workflow that can be applied to similar rare disease datasets.

The clinical implications of these findings are significant. Recognition of dominant symptom clusters and tumor-associated phenotypes may improve diagnostic confidence, support individualized surveillance strategies, facilitate earlier identification of high risk patients, and enhance clinical decision making. The observed relationships among cutaneous, neurological, skeletal, and tumor related manifestations also provide a foundation for more personalized risk stratification in NF1 management.

From an artificial intelligence perspective, this study demonstrates how exploratory analytics can inform the development of explainable clinical decision support systems. The proposed multi-layer framework integrates feature engineering, machine learning, explainable AI, and interactive visualization into a unified architecture capable of supporting future predictive modeling. Such systems have the potential to improve transparency, clinician trust, and adoption of AI-assisted healthcare technologies.

Future research should extend this work by incorporating longitudinal clinical data, genomic variants, imaging biomarkers, and patient reported outcomes to enable multimodal precision medicine. Prospective multicenter validation of the proposed framework, together with implementation of advanced predictive algorithms such as Graph Neural Networks, ensemble learning, and explainable AI techniques, will further enhance clinical applicability. Ultimately, integrating comprehensive phenotypic characterization with intelligent decision-support systems may contribute to earlier diagnosis, improved prognostic assessment, and personalized management of patients with Neurofibromatosis Type 1.

References

[1] Genç, A., Ünal, Y., Ceylan, A. C., Yýldýz, M., Çelik, H. (2025). Clinical and molecular characterization of 148 pediatric neurofibromatosis type 1 patients: A single-center study identifying 14 novel variants. European Journal of Pediatrics, 184, 751. https://doi.org/10.1007/s00431-025-06611.

[2] Williams, V. C., Lucas, J., Babcock, M. A., Gutmann, D. H., Korf, B., Maria, B. L. (2009). Neurofibromatosis type 1 revisited. Pediatrics, 123 (1), 124–133. https://doi.org/10.1542/peds.2007-3204.

[3] Gutmann, D. H., Ferner, R. E., Listernick, R. H., Korf, B. R., Wolters, P. L., Johnson, K. J. (2017). Neurofibromatosis type 1. Nature Reviews Disease Primers, 3 (1), 17004. https://doi.org/10.1038/nrdp.2017.4.

[4] Blakeley, J. O., Plotkin, S. R. (2016). Therapeutic advances for the tumors associated with neurofibromatosis type 1, type 2, and schwannomatosis. Neuro-Oncology, 18 (5), 624–638. https://doi.org/10.1093/neuonc/nov200.

[5] Stewart, D. R., Korf, B. R., Nathanson, K. L., Stevenson, D. A., Yohay, K. (2018). Care of adults with neurofibromatosis type 1: A clinical practice resource of the American College of Medical Genetics and Genomics (ACMG). Genetics in Medicine, 20 (7), 671–682. https://doi.org/10.1038/gim.2018.28.

[6] Wilson, B. N., John, A. M., Handler, M. Z., Schwartz, R. A. (2021). Neurofibromatosis type 1: New developments in genetics and treatment. Journal of the American Academy of Dermatology, 84 (6), 1667–1676. https://doi.org/10.1016/j.jaad.2020.07.105.

[7] Diaz, E., Bergqvist, C., Peiffer, B., Ferkal, S., Pasmant, E., Wolkenstein, P. (2023). In hospital clinical features, morbidity, and mortality of patients with neurofibromatosis 1 in France: A nationwide, population based retrospective cohort study. Journal of Investigative Dermatology, 143 (12), 2408–2415.e7. https://doi.org/10.1016/j.jid.2023.04.034.

[8] Kodali, N., Kumar, K. D., Schwartz, R. A. (2024). The role of scoliosis on the comorbidity and demographics of neurofibromatosis type 1 patients: A retrospective analysis of the National Inpatient Sample database. Experimental Dermatology, 33(1), e14996. https://doi.org/10.1111/exd.14996.

[9] Tanito, K., Ota, A., Kamide, R., Nakagawa, H., Niimura, M. (2014). Clinical features of 58 Japanese patients with mosaic neurofibromatosis 1. Journal of Dermatology, 41 (8), 724–728. https://doi.org/10.1111/1346-8138.12567.

[10] Kang, E., Kim, Y. M., Seo, G. H., Kim, G. H., Kim, J. H., Yoo, H. W. (2020). Phenotype categorization of neurofibromatosis type I and correlation to NF1 mutation types. Journal of Human Genetics, 65 (2), 79–89. https://doi.org/10.1038/s10038-019-0695.

[11] National Institutes of Health Consensus Development Conference. (1988). Neurofibromatosis: Conference statement. Archives of Neurology, 45 (5), 575–578.

[12] Hou, Y., Zong, X., Wu, X., Liu, D., Wolters, P. L., Janusz, J., Stavinoha, P. L., Payne, P. R. O. (2023). Developmental patterns and predictors of academic functioning in children with neurofibromatosis type 1: Results from the largest international dataset created by integrative data analysis [Manuscript submitted for publication]. *SSRN*. https://doi.org/10.2139/ssrn.4610014.

[13] Gupta, A., Hillis, E., Oh, I. Y., Morris, S. M., Abrams, Z., Foraker, R. E., Gutmann, D. H., Payne, P. R. O. (2025). Evaluating dimensionality reduction of comorbidities for predictive modeling in individuals with neurofibromatosis type 1. JAMIA Open, 8 (1), ooae157. https://doi.org/10.1093/jamiaopen/ooae157.

[14] Berisha, V., Krantsevich, C., Hahn, P. R., Hahn, S., Dasarathy, G., Turaga, P., & Liss, J. (2021). Digital medicine and the curse of dimensionality. NPJ Digital Medicine, 4, 153. https://doi.org/10.1038/s41746-021-00521-5.

[15] Rasmy, L., Tiryaki, F., Zhou, Y., Li, Y., Xu, H. (2020). Representation of EHR data for predictive modeling: A comparison between UMLS and other terminologies. Journal of the American Medical Informatics Association, 27 (10), 1593–1599. https://doi.org/10.1093/jamia/ocaa149.

[16] Liu, N., Chee, M. L., Koh, Z. X., Chen, Z., Ong, M. E. H. (2021). Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department. BMC Medical Research Methodology, 21, 74. https://doi.org/10.1186/s12874-021-01258-1.

[17] Zöllner, F. G., Emblem, K. E., Schad, L. R. (2012). SVM-based glioma grading: Optimization by feature reduction analysis. Zeitschrift für Medizinische Physik, 22 (3), 205–214. https://doi.org/10.1016/j.zemedi.2011.10.005.

[18] Basthikodi, M., Chaithrashree, M., Ahamed Shafeeq, B. M., & Gurpur, A. P. (2024). Enhancing multiclass brain tumor diagnosis using SVM and innovative feature extraction techniques. Scientific Reports, 4 26023. https://doi.org/10.1038/s41598-024-77046.

[19] Bhatele, K. R., Bhadauria, S. S. (2022). Machine learning application in glioma classification: Review and comparison analysis. Archives of Computational Methods in Engineering, 29, 247–274. https://doi.org/10.1007/s11831-021-09574-5.

[20] Kumar, A., Jha, A. K., Agarwal, J. P., Yadav, M., Badhe, S., Sahay, A., Epari, S., Sahu, A., Bhattacharya, K., Chatterjee, A., Gupta, R. K. (2023). Machine-learning-based radiomics for classifying glioma grade from magnetic resonance images of the brain. Journal of Personalized Medicine, 13(6), 920. https://doi.org/10.3390/jpm13060920.

[21] Hassan, M. F., Al-Zurfi, A. N., Abed, M. H., Ahmed, K. (2024). An effective ensemble learning approach for classification of glioma grades based on novel MRI features. Scientific Reports, 14, 11977. https://doi.org/10.1038/s41598-024-62592-8.

[22] Joo, B., Ahn, S. S., An, C., Han, K., Choi, D., Kim, H., Park, J. E., Kim, H. S., Lee, S. K. (2023). Fully automated radiomics-based machine learning models for multiclass classification of single brain tumors: Glioblastoma, lymphoma, and metastasis. Journal of Neuroradiology, 50(4), 388–395. https://doi.org/10.1016/j.neurad.2022.10.006.

[23] Vidyadharan, S., Rao, B. V. V. S. N. P., Yogeeswari, P., Kesavadas, C., Rajagopalan, V. (2024). Accurate low and high grade glioma classification using free water eliminated diffusion tensor metrics and ensemble machine learning. Scientific Reports, 14, 19844. https://doi.org/10.1038/s41598-024-70774-1.

[24] Dorfner, F. J., Patel, J. B., Kalpathy-Cramer, J., Gerstner, E. R., Bridge, C. P. (2025). A review of deep learning for brain tumor analysis in MRI. NPJ Precision Oncology, 9, 2. https://doi.org/10.1038/s41698-024-00725-0.

[25] Mohamed Musthafa, M., Mahesh, T. R., Vinoth Kumar, V., Guluwadi, S. (2024). Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. BMC Medical Imaging, 24, 107. https://doi.org/10.1186/s12880-024-01267-0.

[26] Farschtschi, S., Mautner, V. F., McLean, A. C. L., Schulz, A., Friedrich, R. E., Rosahl, S. K. (2020). The neurofibromatoses. Deutsches Ärzteblatt International, 117 (20), 354–360. https://doi.org/10.3238/arztebl.-2020.0354.

[27] Sharafi, P., Arslan, H., Ersoy, S., Varan, A., Ayter, [a]. (2025). Neurofibromatosis Type 1; Clinical symptoms of familial and sporadic cases [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.5505/TurkH-ijyen.2025.06337.