Machine Learning-Based Screening of Narcolepsy Type 1 Using Cataplexy Questionnaires and HLA Biomarkers.

Giorgio Ricciardiello Mejia [1], Andreas Brink-Kjaer[2] , Emmanuel Mignot [1]

[1]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Palo Alto, CA USA

[2]Department of Health Technologies, Technical University of Denmark, Kongens Lyngby, Denmark

Stanford
Psychiatry and Behavioral Sciences

Background

Narcolepsy Type 1 (**NT1**) is a **chronic neurological disorder** caused by hypocretin deficiency and characterized by excessive daytime sleepiness and **cataplexy**—a **sudden**, **emotion-triggered loss of muscle tone** that is highly specific to NT1. While **gold-standard diagnostics** such as the Multiple Sleep Latency Test (MSLT) and cerebrospinal fluid (CSF) hypocretin measurements exist, they are **costly**, **invasive**, and often **unavailable**. Additionally, the **HLA-DQB1*06:02 allele—strongly associated with NT1**—is present in up to 100% of patients but also in ~25% of the general population, limiting its standalone diagnostic value.

Due to **NT1's low prevalence (~30 per 100,000 adults)** and the complexity of diagnosis, **many patients remain undiagnosed or misclassified for years**, delaying access to effective treatments. This study proposes a scalable screening tool combining a validated cataplexy questionnaire and HLA typing, aiming to improve early NT1 identification **using machine learning** (ML) methods. By capturing the diverse emotional and physical triggers of cataplexy, our ML approach seeks to **provide a fast, cost effective, and robust screening tool before referral for confirmatory testing.**

Methods

We analyzed subjective responses from the Stanford Sleep Inventory (SSI) cataplexy questionnaire of **1,207 adult patients** drawn from two studies on narcolepsy with cataplexy (NT1) within the SNC (Anic-Labat et al., 1999 [2]; Okun et al., 2002 [1]). The study sample included **280 patients with NT1** of whom **91.8% (257/280) tested positive for the HLA-DQB10602** allele. The **control group** comprised of **901 adult** patients with primary sleep disorders—such as obstructive sleep apnea, idiopathic hypersomnia, and insomnia—or no cataplexy diagnosis. **NT1 diagnosis** was based on **ICSD-3 definition**.

Descriptive and inferential statistics were used to compare demographic, clinical, and questionnaire-derived variables between NT1 cases and controls. **Continuous variables** were reported as **means ± standard deviations (SD) along with the number of observations**, whereas **binary** variables were reported as **percentages of positive responses with raw counts (n/N)**. For continuous variables, group comparisons were performed using the **Mann–Whitney U test**, and the **rank-biserial correlation** ($r_{rb} = \frac{z}{\sqrt{n_1+n_2}}$) was reported as the effect size where Z is the standardizes test statistic and n_1 and n_2 are group sizes. For binary variables **contingency tables** were analyzed using **Fisher's exact** test (expected cell count < 5) or **Chi-square** test otherwise. Odds ratios ($\frac{a}{b}\frac{d}{c}$) were reported as the effect size, with a, b, c, and d indicating the cell counts form the contingency table. To control for Type I error, **Bnehamini-Hochberg** procedure was implemented to adjust p-values and controls for **false discovery rate**. Both uncorrected and adjusted p-values are reported.

A total of **27 features** were included in the analysis, of which 26 were binary and one was continuous (ESS). The binary variables captured cataplexy emotional triggers, including anger, laughter, joking, quick verbal response, romantic moments, elation, and playing intense games, as well as cataplexy-related muscle weakness in specific locations such as the speech muscles, head, jaw, hands, and knees (**Figure 1**)**.** To ensure model invariance to **demographic factors, age, gender and race were excluded from the feature set**.

Four different feature sets were tested in the study, varying based on the number of features (k) and the inclusion of DQB1*06:02 typing. The **full questionnaire** was evaluated both **with (k=27) and without DQB1*06:02 allele (k=26)**. Additionally, a **reduced questionnaire**, consisting of **key cataplexy features** mapping muscle weaknesses with cataplexy triggers (**Figure 1**), was tested both with **(k =11)** and without DQB1*06:02 allele **(k =10)**.

Five machine learning models were selected: Support Vector Machine (SVM, Lasso, Elastic Net , Linear Discriminant Analysis, and XGBoos. **All models were trained using the entire training data set using the same stratified 5-fold cross-validation**. The **veto rule** was applied to **reduce false-positive** NT1 classifications by correcting predictions for individuals who were HLA-DQB1*06:02 negative, as NT1 cases are always HLA positive. Specifically, for each model and validation fold, predictions labeled as false positives (FP) were overwritten as NT1 negative if the patient was DQB1*06:02 negative. Models were compared to the Okun fuzzy logic previously proposed [1]

Variable	Cases	Controls	Effect Size	p-value uncorrected	p-value FDR corrected
Age$^{\beta}$	45.85±16.82 (279)	48.90±14.78 (914)	0.061	3.53×10^{-02}	4.05×10^{-02}
BMI$^{\beta}$	28.14±6.00 (233)	28.83±7.12 (716)	0.017	5.93×10^{-01}	6.25×10^{-01}
Gender (Male) (yes)$^{\alpha}$	48.2% (1350/280)	66.9% (618/924)	0.461	2.36×10^{-08}	3.40×10^{-08}
Age sleep complaints$^{\beta}$	20.55±9.98 (170)	37.85±16.93 (583)	0.436	5.60×10^{-33}	9.93×10^{-33}
ESS Score$^{\beta}$	**18.74±3.23 (277)**	**11.19±6.07 (926)**	**-0.511**	$\mathbf{1.50\times10^{-70}}$	$\mathbf{3.25\times10^{-70}}$
Naps$^{\beta}$	9.77±8.41 (166)	-			
Disturbed nocturnal sleep (yes)$^{\alpha}$	81.5% (141/173)	66.8% (599/897)	2.192	2.00×10^{-04}	2.00×10^{-04}
Sleep paralysis (yes)$^{\alpha}$	73.4% (179/244)	-			
Sleep paralysis age onset	20.94±9.85 (85)	-			
Hallucinations (yes)$^{\alpha}$	74.7% (1830/245)	-			
Hallucinations Age Onset	19.29±10.91 (94)	-			
MSLT Age	34.02±15.50 (63)	-			
MSLT$^{\beta}$	2.03±1.62 (279)	-			
REM latency$^{\beta}$	36.63±45.20 (135)	-			
Sleep latency$^{\beta}$	82.67±10.61 (135)	-			
Cataplexy medication (yes)$^{\alpha}$	34.8% (810/233)	17.6% (60/34)	2.487	7.29×10^{-02}	8.12×10^{-02}
HLA-DQB1*06:02 (yes)$^{\alpha}$	91.8% (257/280)	24.6% (2280/927)	34.257	3.15×10^{-89}	1.02×10^{-88}
After athletic activities (yes)$^{\alpha}$	24.6% (69/280)	23.3% (216/927)	1.076	7.02×10^{-01}	7.02×10^{-01}
Angry (yes)$^{\alpha}$	**68.6% (192/280)**	**7.0% (65/927)**	**28.934**	$\mathbf{5.79\times10^{-10}}$	$\mathbf{2.82\times10^{-10}}$
Discipline children (yes)$^{\alpha}$	49.6% (139/280)	2.7% (25/927)	35.568	6.50×10^{-89}	1.95×10^{-88}
During sexual intercourse (yes)$^{\alpha}$	18.2% (51/280)	8.0% (74/927)	2.567	1.49×10^{-06}	2.08×10^{-06}
Embarrassed (yes)$^{\alpha}$	32.1% (90/280)	4.5% (42/927)	9.981	7.09×10^{-38}	1.38×10^{-37}
Hear or tell a joke (yes)$^{\alpha}$	**80.7% (226/280)**	**2.0% (19/927)**	**200.008**	$\mathbf{7.67\times10^{-18}}$	$\mathbf{1.49\times10^{-17}}$
Laugh (yes)$^{\alpha}$	**96.8% (271/280)**	**4.9% (45/927)**	**590.178**	$\mathbf{1.76\times10^{-20}}$	$\mathbf{6.85\times10^{-20}}$
Moved by something emotional (yes)$^{\alpha}$	59.3% (166/280)	4.6% (43/927)	29.936	1.01×10^{-98}	3.94×10^{-98}
Quick response cataplexy (yes)$^{\alpha}$	67.1% (188/280)	2.9% (27/927)	68.116	7.43×10^{-13}	7.24×10^{-13}
Remember an emotional moment (yes)$^{\alpha}$	47.1% (132/280)	3.6% (33/927)	24.162	1.88×10^{-76}	4.89×10^{-76}
Startled (yes)$^{\alpha}$	50.7% (142/280)	9.5% (88/927)	9.81	7.11×10^{-53}	1.46×10^{-52}
Tense (yes)$^{\alpha}$	33.9% (95/280)	9.6% (89/927)	4.835	8.36×10^{-23}	1.36×10^{-22}
MW age onset$^{\beta}$	24.37±11.98 (255)	29.42±15.33 (247)	0.152	6.00×10^{-04}	8.00×10^{-04}
MW head and shoulder dropping (yes)$^{\alpha}$	**85.4% (239/280)**	**6.8% (63/927)**	**79.944**	$\mathbf{5.78\times10^{-15}}$	$\mathbf{7.52\times10^{-15}}$
MW in hand and arms (yes)$^{\alpha}$	82.1% (230/280)	16.5% (153/927)	23.271	2.35×10^{-94}	8.35×10^{-94}
MW jaw sagging (yes)$^{\alpha}$	77.9% (218/280)	8.3% (77/927)	38.814	1.06×10^{-12}	6.91×10^{-12}
MW legs and knees (yes)$^{\alpha}$	**97.5% (273/280)**	**35.8% (332/927)**	**69.895**	$\mathbf{1.27\times10^{-72}}$	$\mathbf{3.10\times10^{-72}}$
MW speech becomes slurred (yes)$^{\alpha}$	**63.6% (178/280)**	**2.0% (19/927)**	**83.397**	$\mathbf{1.20\times10^{-13}}$	$\mathbf{9.40\times10^{-13}}$

Table 1 *Distribution of demographic, clinical, emotional triggers for NT1, muscle weakness for NT1, and laboratory test variables stratified by cases (individuals with narcolepsy type 1 [NT1] as defined by ICSD-3 criteria) and controls. Continuous variables (denoted by subscript β) are reported as mean ± standard deviation (number of observations), and binary variables (denoted by subscript α) as percentage of positive responses (n/N). Statistical comparisons were performed using the Mann-Whitney U test for continuous variables and Chi square test for categorical variables. Effect sizes are reported rank-biserial correlation (for continuous variables) and odds ratios (for binary variables). P-values are shown both uncorrected and corrected for multiple comparisons using the Benjamini-Hochberg false discovery rate (FDR) procedure. For comparison with less than 10 samples on either group, first order statistics are reported. ESS – Epworth Sleepiness Scale, BMI – Body Mass Index, MW – Muscle weakness.*

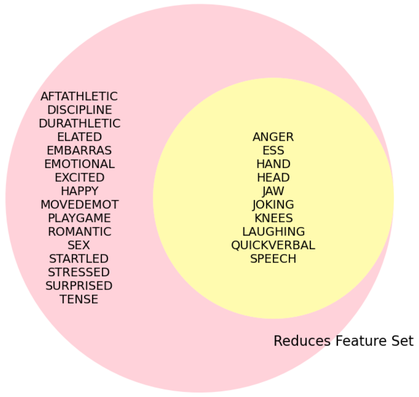

Figure 1 *Ven diagram of the shared features between the full feature (k=26) set and reduced feature set (k=10) without the inclusion of the DQB1*06:02 allele. The reduced feature set contains Anger, ESS, Hand, Jaw, Joking, Kness, Laughing, Quickverbal, and Speech as emotion responses and location for muscle weaknes events.*

Figure 2 Evaluation metrics on the averaged validation fold of the proposed models under a stratified 5-fold cross-validation. The models compared include XGBoost, Elastic Net, SVM, Lasso, logistic regression variants, Okun Decision Tree, LDA, and ESS optimal thresholding. The classification target is NT1 vs. controls. Among the models, Elastic Net demonstrated the best balance between specificity and sensitivity across all feature set configurations, achieving consistently high specificity (≥0.98) while maintaining strong sensitivity (≥0.96). Prevalence adjusted PPV and PPV apparent are presented.

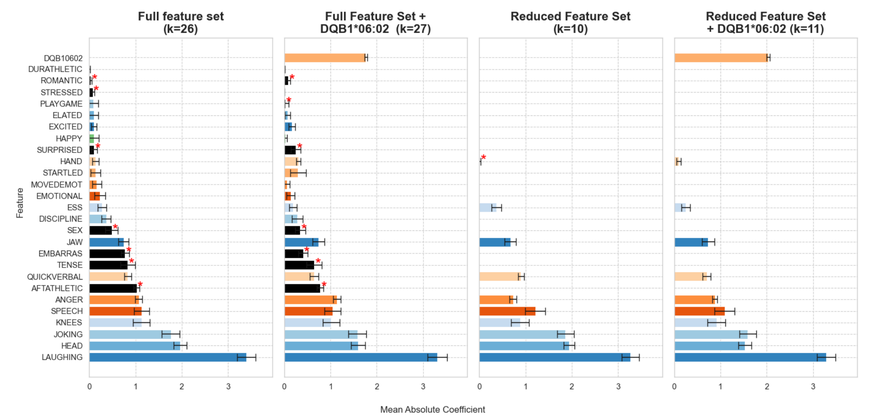

Figure 3 Feature importance from the Elastic Net model across different feature sets. The x-axis represents the mean absolute coefficient, indicating the predictive strength of each feature. The analysis compares a full feature set and a reduced feature set, both with and without HLA-DQB1*06:02. Features with negative coefficients are represented by black bars with red asterisks, showing an inverse relationship with NT1 classification. Standard errors are computed from the average of the coefficient across all the training folds.

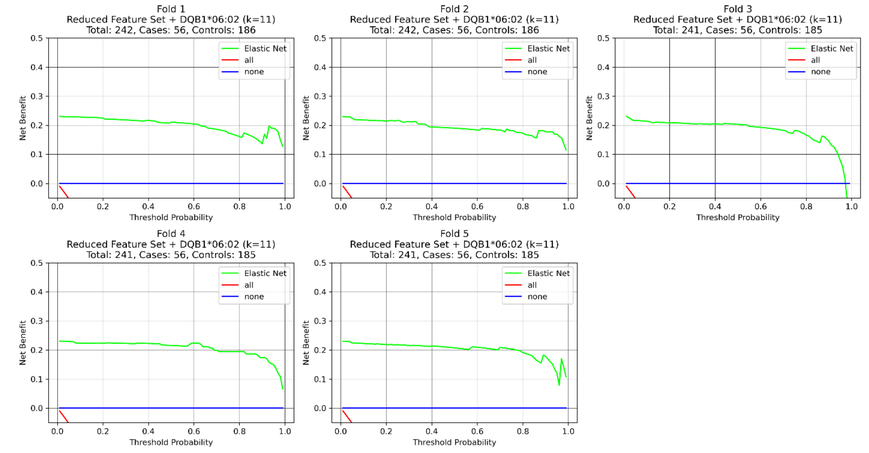

Figure 4 Each subplot in the figure presents the decision curves for each validation fold (5) of the Elastic Net model utilizing the Reduced Feature Set + HLA-DQB106:02 typing. The net benefit, plotted against varying threshold probabilities, is compared to the "treat-all" (red) and "treat-none" (blue) strategies. Curves above these benchmarks indicate improved clinical utility through more accurate identification of NT1 cases and fewer unnecessary interventions. The decision curves aids on the selection of the probability threshold to implement in the classifier for distinguishing cases and controls.

This work was supported by the ResMed Foundation

Results

Elastic Net consistently outperformed tree-based models like XGBoost across all feature with a better trade-off between specificity and sensitivity. Using the reduced (k=11) with HLA typing, Elastic Net achieved a **specificity of 0.981 (95% CI: 0.9763–0.9871)** and **sensitivity of 0.942 (95% CI: 0.9055–0.9802)**, with an F1 score of 0.9409. The PPV, adjusted for NT1 population prevalence, was 0.0166, while the apparent PPV within the sample was 0.94 as depicted in **Figure 2.**

When excluding HLA typing within the reduced feature set (k=10), Elastic Net maintained strong performance with **specificity 0.976 (95% CI: 0.9699–0.9826)**, **sensitivity 0.917 (95% CI: 0.8786–0.9571)**, and F1 score 0.9191. The adjusted PPV was 0.0123 and the apparent PPV was 0.922. With the **full feature set (k=27)** and HLA typing, performance slightly improved with **specificity 0.983 (95% CI: 0.9771–0.9905)**, **sensitivity 0.935 (95% CI: 0.8997–0.9717)**, F1 score 0.9405, PPV 0.0227, and apparent PPV 0.9464. The rule-based ESS threshold model derived from maximizing the Youden's index resulted in an ESS of 15 as threshold for classification.

Figure 3 depicts the mean absolute coefficients from the Elastic Net model, which reflect each feature's contribution to the classification of NT1. Across all feature sets, **laughing, head, joking,** and **knees** consistently ranked among the **strongest predictors**. The inclusion of HLA-DQB1*06:02 notably increased the predictive weight of key features. Negative coefficients e.g., "DISCIPLINE" or "ROMANTIC" were negatively correlated with NT1. The **reduced feature set preserved much of the predictive power of the full set**, highlighting the robustness of a minimal yet highly discriminative questionnaire. The veto rule further improved performance when HLA was absent in the feature set. Remarkably, the biggest improvement was in the PPV.

To evaluate **clinical utility**, **decision curve analysis** was conducted for the Elastic Net. **Figure 4** depicts the decision curve for each validation fold using the reduced feature set (k=11). The model obtained **consistently a positive net benefit** across a wide range of threshold probabilities in all five validation folds. In each subplot, the Elastic Net curve (green) remained well above both the "treat-all" (red) and "treat-none" (blue) strategies, indicating superior performance in correctly identifying NT1 cases while minimizing false positives.

Conclusion

Findings of this study demonstrate that **machine learning models**, particularly Elastic Net, provide a **highly specific and scalable approach for screening NT1** in the general population. By leveraging a brief, validated cataplexy questionnaire and **incorporating HLA-DQB1*06:02 typing**, the model **achieved high discriminative performance**, with specificity up to 0.99 and sensitivity up to 0.98 across validation folds. The addition of HLA-DQB1*06:02 increased predictive precision by reducing false positives, and the **post hoc veto rule offered an effective correction mechanism when genetic data were unavailable**. These findings support the clinical utility of a non-invasive, questionnaire-based ML tool as a screening tool prior to confirmatory testing. Such an approach could **reduce unnecessary referrals for the MSLT**, **making early NT1 identification more accessible, cost-effective, and feasible at scale.**

References

[1] M. L. Okun, L. Lin, Z. Pelin, S. Hong, and E. Mignot, "Clinical aspects of narcolepsy-cataplexy across ethnic groups," *Sleep*, vol. 25, no. 1, pp. 27–35, 2002.

[2] S. Anic-Labat, C. Guilleminault, H. C. Kraemer, J. Meehan, J. Arrigoni, and E. Mignot, "Validation of a cataplexy questionnaire in 983 sleep-disorders patients," *Sleep*, vol. 22, no. 1, pp. 77–87, 1999.

Mignot Laboratory

GitHub