



OPEN Prediction of recovery grade in Bell's palsy using machine learning models

Sang-Yoon Han^{1,8}, Jeong Rye Park^{2,8}, Heonjeong Oh³, Sung-Min Park⁴, Jongyook Park^{5,9}✉ & Young Ho Kim^{6,7,9}✉

While most cases of Bell's palsy resolve successfully, some patients do not fully recover despite medical treatment. Various tools have been developed to assess the status of facial nerve palsy (FNP), enhance prognostic accuracy, and guide early rehabilitation or more intensive therapies. However, existing evaluation tools often struggle with accurately predicting moderate cases of FNP. Recently, machine learning algorithms have shown promise in predicting clinical outcomes across various diseases by analyzing historical medical data. This study aimed to establish machine learning models for predicting acceptable recovery and House–Brackmann (H–B) grades at each visit using Support Vector Machine (SVM) and Long Short-Term Memory (LSTM). By leveraging data from serial H–B grades for each clinic visit and other clinically significant features, we demonstrated that changes in FNP grades can be predicted with high accuracy, achieving 0.903 accuracy for forecasting acceptable recovery and a Mean Absolute Error (MAE) of approximately 0.46 per H–B grade for each visit. Given the substantial social and emotional impact of FNP, early intervention based on prognosis is crucial. Our machine learning model can aid clinicians in predicting outcomes and providing appropriate consultation and treatment for FNP patients.

Facial nerve palsy (FNP) is classified as either central or peripheral, depending on the suspected primary site of the disease. Central FNP results from upper motor neuron injuries affecting structures such as the cerebral cortex, pons, and corticobulbar tract^{1,2}. In contrast, peripheral FNP occurs due to facial nerve injury, causing paralysis on the ipsilateral side of the face, including the forehead².

Peripheral FNP has diverse causes, including idiopathic factors, varicella-zoster virus infection, chronic otitis media, and trauma. The most common cause is idiopathic, referred to as Bell's palsy. A study reported that the monthly prevalence of Bell's palsy ranges from 7.7 to 9.1 per 100,000³. Since viral infection and inflammation are considered potential causes of Bell's palsy, steroids are widely used for treatment^{1,4}. Although most cases of Bell's palsy resolve successfully, some patients experience incomplete recovery despite medical therapy^{1,4,5}. Surgical treatment may be considered in cases of severe FNP with minimal likelihood of recovery, even after medical management^{1,4,6}. However, its efficacy remains controversial, and it carries risks of postoperative complications and residual sequelae^{6,7}.

Various evaluation tools have been developed to assess the current status of FNP, improve prognostic prediction, and guide early rehabilitation or more intensive treatment. To date, more than 10 grading systems, including the House–Brackmann facial paralysis scale (H–B grade)⁸ and the Sunnybrook facial grading system, have been introduced. The H–B grade system, developed in 1985, is the most widely used⁸. Electrophysiologic evaluations, including electroneurography (ENoG) and electromyography (EMG), have been used for assessment and prognostic evaluation. However, most tests have limited prognostic value due to poor reliability⁹. Notably, ENoG is considered a valuable tool for determining the need for surgical intervention in Bell's palsy and

¹Department of Otolaryngology-Head and Neck Surgery, College of Medicine, Hanyang University, Seoul, Republic of Korea. ²Department of Mathematical Sciences, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea. ³Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Hospital, Seoul, Republic of Korea. ⁴Department of Otorhinolaryngology-Head and Neck Surgery, College of Medicine, Kangnam Sacred Heart Hospital, Hallym University, Seoul, Republic of Korea. ⁵Department of Mathematics, Kyungpook National University, Daegu, Republic of Korea. ⁶Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University Boramae Medical Center, Seoul National University College of Medicine, 5 Gil 20, Boramae-Road, Dongjak-Gu, Seoul 156-707, Republic of Korea. ⁷Department of Otorhinolaryngology-Head and Neck Surgery, Seoul National University College of Medicine, Seoul, Republic of Korea. ⁸Sang-Yoon Han, Jeong Rye Park have contributed equally to this work as a first author. ⁹Jongyook Park, Young Ho Kim have contributed equally to this work as a corresponding author. ✉email: jongyook@knu.ac.kr; yhkiment@gmail.com

predicting prognosis¹⁰. Despite this, ENoG has limitations in predicting recovery rates or the natural course of moderate FNP¹¹. Given that most recovery occurs within the first 3 weeks and approximately 80% of Bell's palsy cases achieve favorable outcomes^{4,5,12,13}, predicting the prognosis of non-recovered cases at an early stage may support clinical decision-making.

Machine learning algorithms have emerged as promising tools for predicting clinical outcomes across various diseases using historical medical data. A previous study demonstrated the feasibility of predicting hearing recovery with a deep learning model¹⁴. Similarly, studies have demonstrated that Recurrent Neural Network (RNN) models can predict visual field recovery^{15,16}. A recent study introduced a strategy using explainable deep learning approaches with attention-based mechanisms and the Synthetic Minority Oversampling Technique (SMOTE), which exhibited good performance in predicting brain tumors and Alzheimer's disease by addressing data imbalance and improving both model stability and interpretability¹⁷. Efforts have also been made to apply deep learning to FNP, with some studies exploring models for the diagnosis and grading of FNP^{18,19}. However, machine learning models for predicting the prognosis of FNP remain largely unexplored.

Herein, this study aims to predict the early prognosis of FNP using clinical data obtained during the early stages after onset, leveraging a machine learning model.

Materials and methods

Subjects inclusion and data curation

We retrospectively reviewed the electronic medical records of 407 patients who visited the clinic or emergency room for FNP between January 1, 2010, and March 31, 2024. Initially, 390 patients were selected who had FNP grades at the initial visit, follow-up information, age between 10 and 100 years, and available data on sex, diabetes, and hypertension. Among them, 279 patients with records of at least two visits with FNP grade evaluations within 30 days of onset and at least one FNP grade evaluation between 30 and 300 days after onset were selected as the study subjects. The subjects were further categorized to create a main dataset for developing the machine learning models and a validation dataset for model evaluation. In consideration of previous findings on sample size recommendations (approximately 1000 for LSTM and 500 for SVM)^{20–25} and the necessity of applying temporal and external validation to prevent information leakage and to simulate clinical circumstances under which predictions would be made for future patients^{26–28}, January 1, 2024, was set as the cutoff date (Fig. 1). Consequently, the training and test datasets were constructed with 267 subjects, and the validation dataset was constructed with 12 subjects. Furthermore, the dataset for predicting recovery of unrecovered patients within 30 days (persistent FNP) was constructed from data obtained from 151 subjects.

Information on demographic factors, and the medical history of FNP was retrieved from electronic medical records. Since delayed treatment (treatment starting more than 7 days after onset), diabetes, and hypertension have been identified in studies as important risk factors for poor FNP prognosis, these features were also retrieved to construct the machine learning models^{29–33}. This study was approved by the Institutional Review Board of Seoul National University Boramae Medical Center (IRB No. 10-2022-90). Given its retrospective nature, the Institutional Review Board of Seoul National University Boramae Medical Center waived the requirement for informed consent. All methods were carried out in accordance with relevant guidelines and regulations. This study was conducted in accordance with the STROBE statement and the TRIPOD statement.

FNP grading system

The H–B grade system was used to evaluate the grades of FNP at each clinic visit (suppl. Table 1)⁸. In this study, ambiguous H–B grades, described as ranges between adjacent grades in the electronic medical records, were defined by adding 0.5 to the lower grade. A grade of II or better was considered acceptable recovery. The mean H–B grade was calculated by averaging the numerical values of the H–B grades for each patient.

Data preprocessing

Missing values and outliers were excluded during the subject inclusion process prior to data preprocessing. For each patient, d_i and g_i represent the date of the i th clinic visit and the corresponding H–B grade, respectively. d_1 was defined as the day before FNP onset with g_1 assigned a grade of 1, and the first hospital visit date, d_2 , was set as day 0. Overlapping subsequences of three clinic visits within the first 30 days were constructed, such as (d_1, d_2, d_3) , (d_2, d_3, d_4) , ..., (d_{i-2}, d_{i-1}, d_i) , with d_k and g_k representing the time point and grade after 30 days. Each subsequence (d_p, d_{i+1}, d_{i+2}) was used to generate features $(d_p, d_{i+1}, d_{i+2}, d_k, g_p, g_{i+1}, g_{i+2})$ for machine learning models. Regression models used both time and grade after 30 days, whereas classification models used only the final outcome.

To standardize the time points by setting the present time point d_{i+2} to 0, the Selected Assessment Days (SAD1–SAD4) were defined as $d_1 - d_{i+2}$, $d_{i+1} - d_{i+2}$, 0, and $d_k - d_{i+2}$, and the Selected Assessment Grades (SAG1–SAG4) were defined as g_p , g_{i+1} , g_{i+2} , and g_k . Additional features included gender, age, hypertension, diabetes, and delayed treatment (treatment starting more than 7 days after onset).

For the classification task predicting patient recovery, patients were labeled as recovered (1) if g_k was less than or equal to 2, and not recovered (0) if g_k was greater than 2. Eleven features—gender, age, hypertension, diabetes, delayed treatment, SAD1, SAD2, SAD4, SAG1, SAG2, and SAG3—were used to construct the model (Suppl. Fig. 1). For regression models predicting H–B grade at specific time points, the 11 features were used, with two out of hypertension, diabetes, and delayed treatment included along with gender, age, SAD1–SAD4, and SAG1–SAG3 (Suppl. Fig. 1).

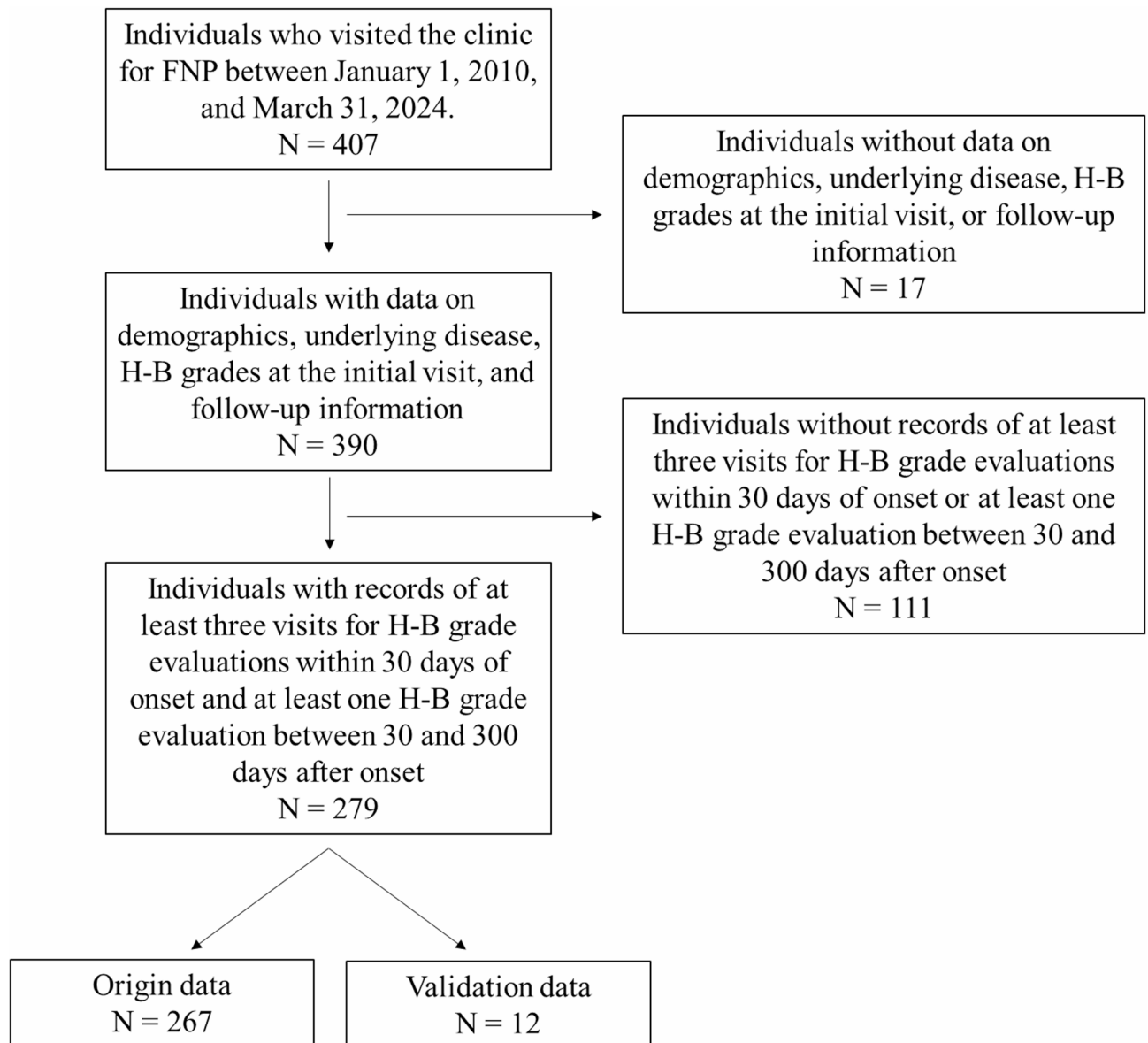


Fig. 1. Subjects inclusion. FNP, facial nerve palsy; H–B grade, House Brackmann grade.

Machine learning models for predicting the prognosis of FNP

Model construction

The predictive models designed and implemented were developed using Python 3.11. For the classification model aimed at predicting patient recovery, scikit-learn 1.5.2 was utilized. For the regression model aimed at predicting the patient’s H–B grade, TensorFlow 2.18.0 and Keras were employed. To ensure reproducibility, all experiments were performed with fixed random seeds (random_state = 42 in scikit-learn and tf.random.set_seed(42) in TensorFlow).

Exploratory analysis of feature importance

To identify the key variables affecting model performance, permutation importance analysis was conducted for all subjects using Support Vector Machine (SVM)³⁴ with an RBF kernel. SAG1 was excluded from this analysis because it had a constant value of 1. All features were standardized to have a mean 0 and a variance of 1 before training. Permutation importance was computed on the entire dataset with accuracy as the metric, repeating 100 random permutations for each feature. This exploratory analysis focused on assessing relative feature influence rather than model generalizability.

Classification model for acceptable recovery of facial nerve palsy

SVM³⁴ was utilized to predict the recovery of FNP patients. Acceptable recovery was defined as grade II or below within 300 days for patients. A total of 668 data points were obtained after preprocessing from 267 subjects. The dataset was split into training (N = 534) and test (N = 134) sets in an 8:2 ratio. An independent sample *t* test was

conducted between the training and test data, confirming that there was no statistically significant difference between the two groups (Suppl. Table 2). All values were normalized to a mean of 0 and a variance of 1. The same analysis was conducted on 393 data from 151 subjects who did not show recovery from FNP within 30 days, with the dataset split into training (N = 314) and test (N = 79) sets in an 8:2 ratio (Suppl. Table 3).

The SVM model was tuned with efficient hyperparameters using the grid search algorithm provided by scikit-learn. Additionally, fivefold cross-validation was performed on the training data to prevent overfitting. In the grid search, the range of the regularization parameter was set from 0.01 to 30, while the range of gamma was set from 0.001 to 10. For the model involving all subjects, parameters were selected based on the highest accuracy.

The selected parameters were $C = 2.7$ and $\gamma = 0.09$. Additionally, the radial basis function (RBF) kernel was used. For the classification models of patients with persistent FNP, the selected parameters were $C = 5.6$ and $\gamma = 0.09$ (Suppl. Table 4). The RBF kernel was also used. The evaluation metrics for the model were accuracy, sensitivity, and F1-score, based on the Confusion Matrix.

Regression model for predicting the recovery course of facial nerve palsy

Long Short-Term Memory (LSTM)^{35,36} was implemented for modeling the prediction of the recovery course of FNP. After preprocessing, 1,253 data were obtained from 267 patients. The dataset was split into training (N = 1002) and test (N = 251) sets in an 8:2 ratio. An independent sample *t* test was conducted between the training and test data, confirming that there was no statistically significant difference between the two groups (Suppl. Table 5). The same analysis was conducted for the 151 patients with persistent FNP, with the data (N = 869) split into training (N = 695) and test (N = 174) sets in an 8:2 ratio. The training and test data showed no statistically significant differences in demographic factors, underlying diseases, clinic visit timing, and the distribution of H–B grades between the two groups (Suppl. Table 6). The validation data consisted of 33 day-grade pairs from 12 individuals and was used to validate the regression model of recovery from FNP for all subjects.

The model consists of three LSTM layers and one output layer. The input data is structured as a 4×3 matrix, representing two past time points, one present time point (SAD3 = 0), and one future time point (SAD4 > 0). Each time point includes the corresponding H–B grade, while the H–B grade of the future time point (the prediction target) is left blank so that the model can learn to generate predictions during training. Additionally, the input includes four key features that are highly related to FNP. These features consist of gender and age, which are essential factors known to influence FNP, along with two additional features selected from delayed treatment, diabetes, and hypertension, and, forming three different feature combinations.

The first LSTM layer consists of 64 units, the second has 32 units, and the third has 16 units. All layers use the tanh activation function, 20% dropout to prevent overfitting, and batch normalization to enhance training stability. The output layer is a Dense layer, designed to produce a single continuous value. Mean Squared Error (MSE) was used as the loss function to minimize the difference between predicted and actual values. Additionally, MSE was also used as a performance evaluation metric for further monitoring. The model was optimized using the Adam optimizer with a learning rate of 0.005.

During training, 10% of the total data was set aside as a validation set to continuously evaluate the model's performance throughout the process. To efficiently derive the optimal model, Early Stopping and ReduceLRonPlateau were applied.

When using all data, the batch size was set to 128. Training was stopped if the validation loss did not improve for 40 epochs, and the learning rate was reduced by 50% if the validation loss did not improve for 20 epochs. For the subjects with persistent FNP, the batch size was adjusted to 64 due to differences in dataset size. In this setting, training was stopped if the validation loss did not improve for 80 epochs, and the learning rate was reduced by 50% if the validation loss did not improve for 30 epochs. To prevent excessive reduction in the learning rate, the minimum learning rate was set to $1e-6$. Detailed model configurations and hyperparameters are provided in Supplementary Table 7. The model maintains the same structure regardless of the feature combination and is designed to perform consistently across all three combinations (Suppl. Fig. 2).

The model was evaluated using Root Mean Squared Error (RMSE), MSE, and Mean Absolute Error (MAE) as performance metrics.

Results

Demographics and clinical characteristics

The mean age of the subjects in the training set was 56.57 ± 17.59 years. There were 136 male subjects (50.94%) and 131 female subjects (49.06%). The mean H–B grade was 3.25 ± 0.84 , with 18.35%, 40.45%, 33.33%, and 7.87% classified as H–B grades II, III, IV, and V, respectively (values represented within the range were classified as severe grades). The mean first visit date after FNP occurrence was 2.35 ± 3.11 days, and the mean follow-up duration was 88.84 ± 49.09 days. Of the subjects, 64.79% received treatment during the first week, while 35.21% started steroid administration 1 week after onset. Among all subjects, 25.09% had diabetes, and 31.09% had hypertension (Table 1).

For the 12 subjects (4 males and 8 females) in the validation set, the mean age was 58.67 ± 19.15 years. Their detailed values for demographic factors, underlying diseases, H–B grades, delayed treatment status, and follow-up patterns are described in Table 1.

Clinical course of facial nerve palsy

The changes in FNP over time were examined and plotted for all 1425 day-grade pairs. FNP was abruptly aggravated during the first week after onset, followed by gradual recovery over time. The most severe grade occurred in the latter half of the first week, with a mean H–B grade of approximately 3.5. By the fifth week, the

Groups	Individuals in the training set (N = 267)	Individuals in the test set (N = 12)
Age (years)	56.57 ± 17.59	58.67 ± 19.15
Gender (M:F)	136 : 131	4 : 8
Diabetes	25.09% (N = 67)	41.67% (N = 5)
Hypertension	31.09% (N = 83)	33.33% (N = 4)
Initial H–B grades	3.25 ± 0.84	3.58 ± 1.18
Grade II	18.35% (N = 49)	16.67% (N = 2)
Grade III	40.45% (N = 108)	25.00% (N = 4)
Grade IV	33.33% (N = 89)	8.33% (N = 2)
Grade V	7.87% (N = 21)	33.33% (N = 4)
Mean first visit date (days)	2.35 ± 3.11	2.25 ± 3.57
Mean follow-up duration (days)	88.84 ± 49.09	67.83 ± 29.27
Treatment		
Early (within 7 days)	64.79% (N = 173)	100.00% (N = 12)
Delayed (after 7 days)	35.21% (N = 94)	0.00% (N = 0)

Table 1. Demographic factors and clinical characteristics of the subjects. M, male; F, female; H–B grade, House Brackmann grade; N, number.

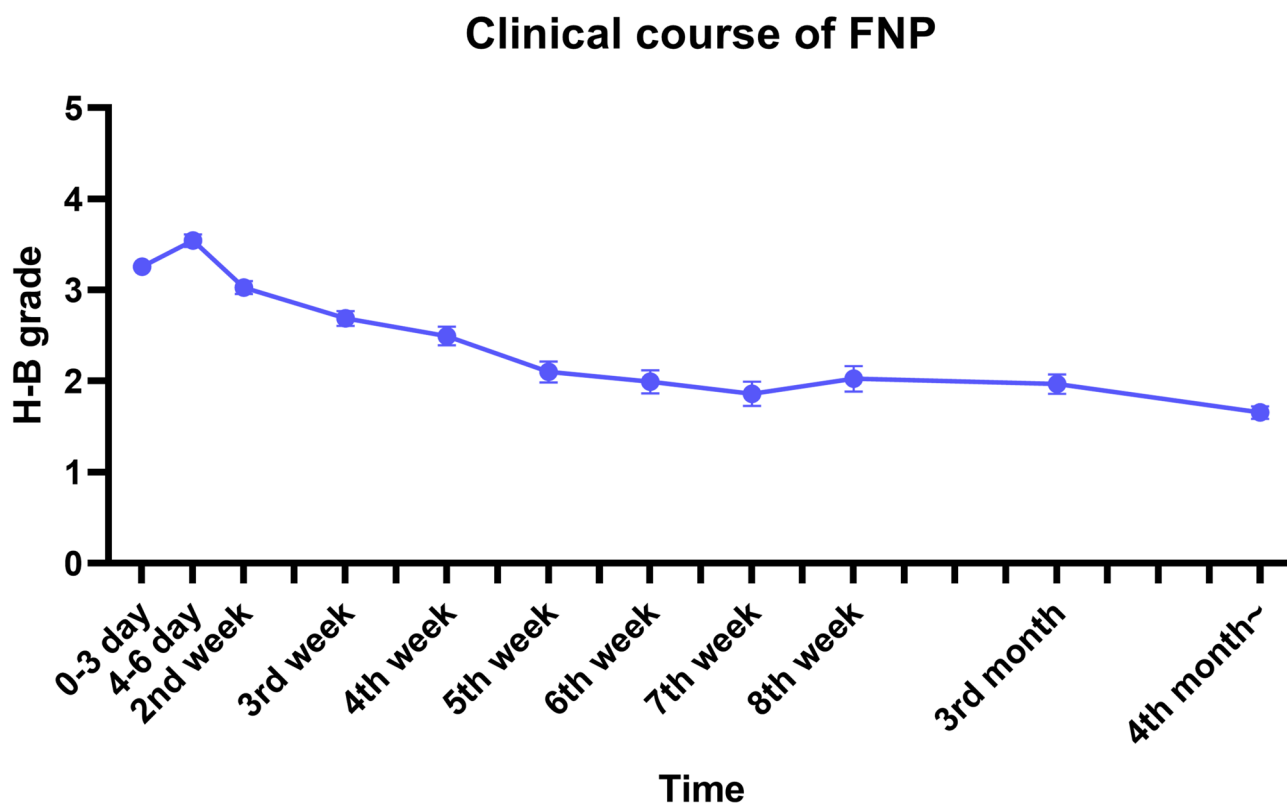


Fig. 2. Clinical recovery course of facial nerve palsy. FNP, facial nerve palsy; H–B grade, House Brackmann grade.

FNP grade had improved to around H–B grade II, which represents an acceptable treatment outcome for FNP (Fig. 2).

Prediction of prognosis in facial nerve palsy using artificial intelligence

Feature importance

Among the features used to construct the models, SAG3 had the highest importance at 24.43%, followed by SAD4 (13.72%). Gender (10.57%), hypertension (10.01%), age (9.57%), and diabetes (7.58%) showed relatively higher values, while the other H–B features exhibited lower importance (Fig. 3).

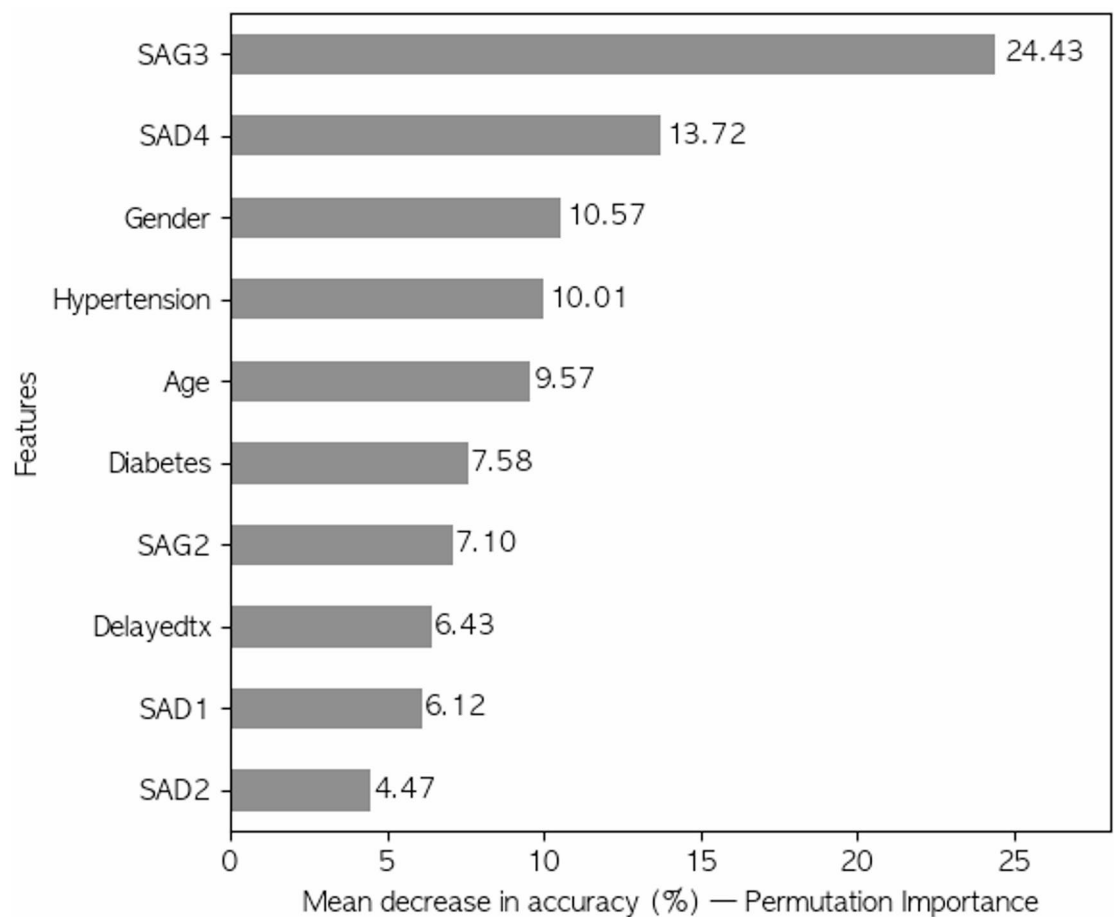


Fig. 3. Feature importance analysis was performed using clinical features. The analysis was conducted using permutation importance for all subjects. For each feature, 100 random permutations were performed. SAD, selected assessment day; SAG, selected assessment grade; Tx, treatment.

Classification models

The classification models, considering H–B grades from two serial visits within 30 days after onset, age, gender, delayed treatment, diabetes, and hypertension, and (Suppl. Table 3), showed an accuracy of 0.903 (95% CI=0.853–0.953), a recall of 0.991 (95% CI=0.970–1.000), and an F1-score of 0.944 (95% CI=0.912–0.975) (Fig. 4).

The same analysis was performed for individuals with persistent FNP. The classification model achieved an accuracy of 0.848 (95% CI=0.769–0.927), a recall of 0.946 (95% CI=0.885–1.000), and an F1-score of 0.898 (95% CI=0.838–0.950) (Fig. 4).

Regression models

H–B grades from two consecutive visits within 30 days after onset, along with age and gender, were used as features in all regression models. Additionally, two of the following variables— delayed treatment, diabetes, and hypertension—were included as features in model 1 (delayed treatment and diabetes), model 2 (delayed treatment and hypertension), and model 3 (diabetes and hypertension). The specific values of each feature are described in Supplementary Table 5. Model 1 had an MAE of 0.478 (95% CI=0.415–0.550) and an MAE of 0.460 (95% CI=0.323–0.606) in the validation data, while Model 2 showed worse MAE values (0.487, 95% CI=0.420–0.559) and MAE values (0.581, 95% CI=0.431–0.760) in the validation data compared to Model 1. Model 3 had an MAE of 0.458 (95% CI=0.401–0.532) and an MAE of 0.481 (95% CI=0.345–0.619) in the validation data (Table 2). The estimated results with Model 3 for the test data were the most accurately predicted, except for H–B grade 5 on SAD4 (Table 2, Fig. 5), showing a similar course with minimal gaps in training loss and validation loss (Suppl. Fig. 2). The estimated course using the validation data also featured a similar progression to the natural course of FNP (Fig. 6) with a low MAE (Table 2).

For individuals with persistent FNP, the same analysis was conducted. The MAE of Model 1, Model 2, and Model 3 were 0.608 (95% CI=0.515–0.701), 0.680 (0.591–0.777), and 0.637 (0.533–0.747) in the test data (Table 2).

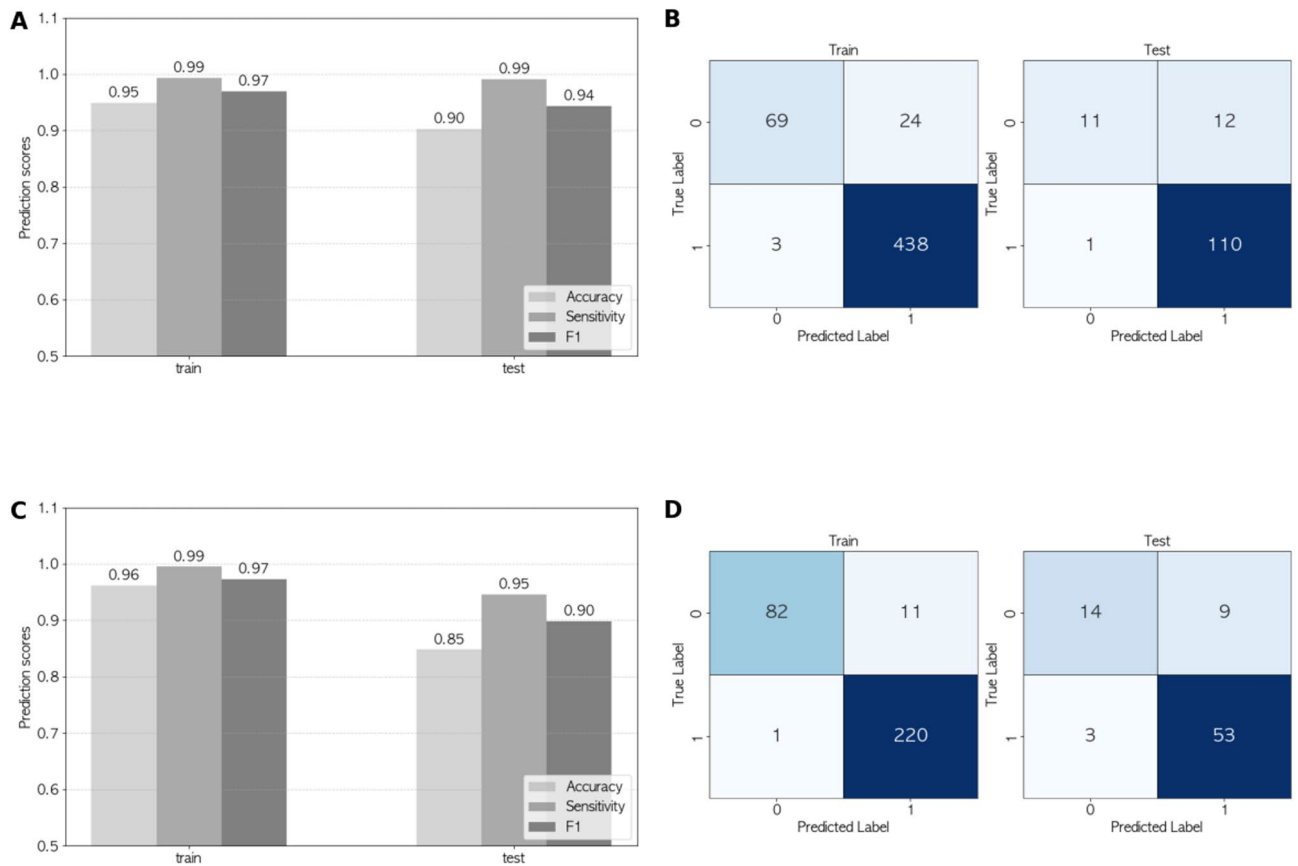


Fig. 4. Prediction scores (A) and confusion matrices (B) of classification models for predicting acceptable recovery from facial nerve palsy. Prediction scores (C), and (D) Confusion matrix for those who had not recovered from facial nerve palsy until 30 days after the initial visit.

Subjects	Models	Model 1		Model 2		Model 3	
		Gender, age, delayed Tx, DM (95% CI)	Gender, age, delayed Tx, DM (95% CI)	Gender, age, delayed Tx, HTN (95% CI)	Gender, age, delayed Tx, HTN (95% CI)	Gender, age, DM, HTN (95% CI)	Gender, age, DM, HTN (95% CI)
All subjects	Dataset	Test (N = 251)	Validation (N = 33)	Test (N = 251)	Validation (N = 33)	Test (N = 251)	Validation (N = 33)
	RMSE	0.720 (0.600–0.848)	0.626 (0.467–0.769)	0.737 (0.611–0.870)	0.753 (0.575–0.925)	0.709 (0.584–0.844)	0.632 (0.464–0.778)
	MSE	0.519 (0.358–0.708)	0.392 (0.223–0.585)	0.544 (0.368–0.737)	0.566 (0.342–0.827)	0.503 (0.332–0.744)	0.399 (0.212–0.582)
	MAE	0.478 (0.415–0.550)	0.460 (0.323–0.606)	0.487 (0.420–0.559)	0.581 (0.431–0.760)	0.458 (0.401–0.532)	0.481 (0.345–0.619)
	P value*	0.054	0.999	0.191	0.113	0.809	0.621
Persistent FNP after 30 days	Dataset	Test (N = 174)	–	Test (N = 174)	–	Test (N = 174)	–
	RMSE	0.892 (0.734–1.055)	–	0.936 (0.796–1.087)	–	0.952 (0.786–1.118)	–
	MSE	0.796 (0.529–1.103)	–	0.877 (0.603–1.188)	–	0.906 (0.616–1.247)	–
	MAE	0.608 (0.515–0.701)	–	0.680 (0.591–0.777)	–	0.637 (0.533–0.747)	–
	P value*	0.675	–	0.779	–	0.144	–

Table 2. Performance comparison of regression deep learning models based on features. Tx, treatment; DM, diabetes mellitus; HTN, hypertension; N, number; RMSE, Root Mean Squared Error; MSE, Mean Squared Error; MAE; Mean Absolute Error; *, the statistical results of *p* values from the paired *t* test between actual House–Brackmann grades and predicted House–Brackmann grades.

Discussion

The present study demonstrated that changes in H–B grades in FNP patients after initial recovery could be successfully estimated using the regression and classification models designed based on SVM and LSTM. The classification models showed good accuracy for all patients. In addition, as shown in Figs. 4 and 5, our machine learning models accurately predicted the clinical course of FNP, showing favorable MAE values for all patients, including those who had not recovered by the first month. Based on these models, grades at specific time points can be predicted, and the expected recovery rates can be derived from three consecutive evaluations of grades.

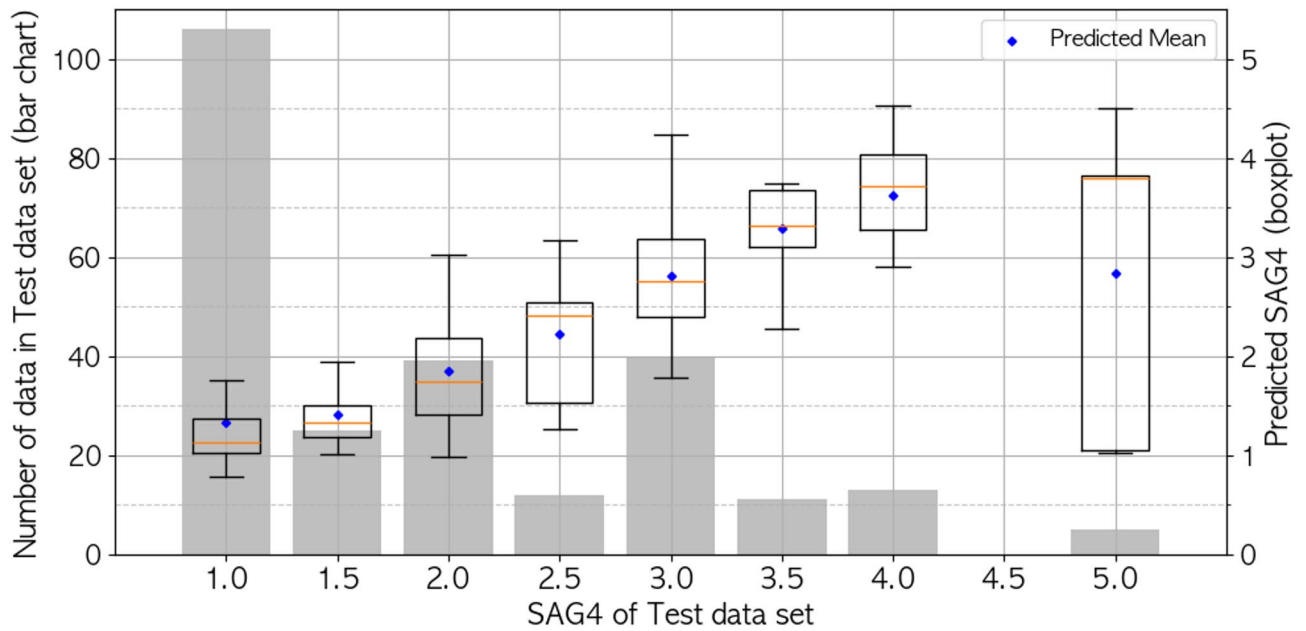


Fig. 5. Predicted values of the House–Brackmann grade on the fourth selected assessment day using regression models constructed with age, gender, diabetes, and hypertension in the test set. The number of test data sets was documented in gray bar. The predicted mean H–B grade is represented by a blue dot, while values exceeding two standard deviations are marked with black circles. SAG, selected assessment grade.

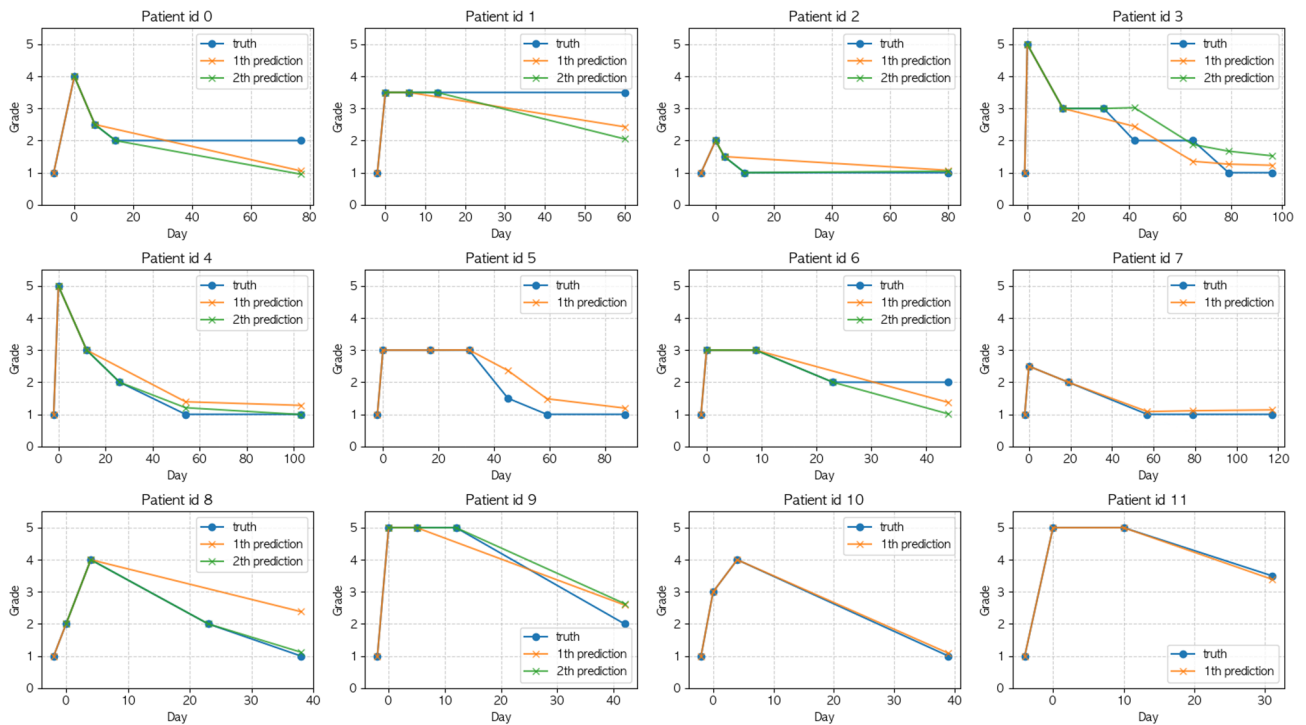


Fig. 6. The predicted recovery course from facial nerve palsy using long short-term memory using features including age, gender, hypertension and diabetes in validation dataset.

This study proposed machine learning models that serve as potential decision-support tools by estimating the recovery course of FNP, thereby assisting clinicians in facilitating early rehabilitation or cosmetic interventions for facial reanimation and in guiding the determination of aggressive treatment plans, such as facial nerve decompression⁴. Considering that the majority of FNP recovers within the first 3 weeks, most patients do not receive rehabilitation. In contrast, patients who do not fully recover tend to be lost to follow-up loss and seek

alternative treatments, such as acupuncture and moxibustion⁴. Given that appropriate rehabilitation for FNP can improve the recovery rate, providing consultation and rehabilitation based on predicted recovery potential is crucial for patients with FNP^{37,38}. Additionally, FNP can result in high socio-emotional costs, and consultation for reconstructive intervention should be provided to patients predicted to have an unfavorable prognosis⁴. However, existing evaluation tools for FNP have limitations in predicting the recovery course, which may delay appropriate interventions⁹. In this regard, the results of this study may be helpful for clinicians in providing appropriate consultation based on the expected changes and recovery of FNP.

The accuracy of the classification model was 0.903 for all subjects, with a recall of 0.991. Given that other estimation tools have poor reliability and ENoG is limited in predicting the prognosis for mild or severe FNP^{9,11}, our models can predict with better accuracy than conventional tools. Surprisingly, the recovery estimation was successfully maintained with an accuracy of 0.848 and a recall of 0.946 for patients with persistent FNP. Furthermore, the prediction method was developed using serial H–B grades from three clinic visits and did not require additional examinations or discomfiting procedures. Using the classification model, physicians can predict successful recovery without additional time, cost, or procedures, while maintaining good prediction rates.

Prediction of the patient's H–B grade is a regression problem, where the goal is to predict the exact H–B grade at a specific time point. Since a patient's condition changes over time, it is important to track how it develops. The MAE of our regression models was about 0.458 after adding diabetes, hypertension, age, and gender as features in the machine learning model. Since the grade consists of ordinal values, an MAE of 0.458 in model 3 may indicate that the predicted values are similar to the actual values. Furthermore, those with persistent FNP also showed good prediction for H–B grades at the fourth visit, with a MAE of approximately 0.61 when considering demographics, diabetes, and delayed treatment. Based on the regression model, the estimated natural course of FNP can be outlined. Therefore, the classification model and regression model of this study can be helpful in explaining the natural course of personalized recovery, reassuring patients, and consulting on further treatment.

According to the clinical course of FNP described in Fig. 2, the mean values of the H–B grades showed successful recovery about 1 month after onset. However, the present study revealed that some patients did not recover within 1 month after the initial visit. Surprisingly, our machine learning model also showed good predictions for these patients, in terms of grade changes and acceptable recovery rates. Based on the machine learning models from our study, clinicians may benefit from more accurate predictions of facial palsy and provide appropriate interventions for patients with FNP without the need for complex evaluations.

Some previous studies identified diabetes and hypertension as risk factors for FNP and found them to be associated with poor prognosis^{29–31}. Delayed treatment was also a significant risk factor for the poor prognosis of FNP^{32,33}. Therefore, considering these factors when predicting the prognosis of FNP may improve classification accuracy, contributing to the higher accuracy observed in our results. Based on the results of the present study, LSTM incorporating diabetes and hypertension performed best for all subjects, which was consistent with the findings from permutation importance, while the inclusion of delayed treatment and diabetes as features yielded the highest accuracy for subjects with persistent FNP. However, our study showed inconsistent results between the training set and the validation data for all subjects. This might be due to delayed treatment not being present in the validation data. Further studies with larger validation data sets may provide results with higher accuracy and greater consistency with the training set.

The findings from this study demonstrated that SAG4 was most significantly associated with SAG3, followed by SAD4. As shown in Fig. 2, the typical natural course of FNP remains relatively stable over time, exhibiting a regression pattern. Therefore, the grade from the most recent evaluation and the date of evaluation are likely to have a significant impact on predicting a given point³⁹.

In addition, age also showed a higher association with SAG4. Consistent with the present study, previous research showed a relationship between facial palsy recovery and age⁵. With increasing age, neural regeneration becomes limited⁴⁰, endothelial function and capillary density decline^{41,42}, low-grade inflammation persists⁴³, muscle atrophy accelerates⁴⁴, recovery of neuromuscular junctions diminishes⁴⁴, and denervation atrophy progresses more rapidly⁴⁴, all of which may contribute to poorer facial nerve recovery in older individuals. Therefore, as suggested in the present study, poorer outcomes in older individuals should be anticipated by taking additional prognostic factors into account, and prompt, diverse therapeutic strategies should be implemented accordingly.

In the present study, gender was considered a potential factor for predicting FNP recovery. However, its role as a prognostic indicator remains controversial in previous studies^{45–48}. One study reported that in women, earlier treatment and younger age were associated with better recovery⁴⁷, whereas other studies found no significant association^{45,46,48}. Evidence from animal and cellular studies suggests that females may exhibit faster or more effective nerve regeneration than males due to a combination of factors^{49–52}. Enhanced Schwann cell activity, faster remyelination, a supportive distal nerve microenvironment, and estrogen contribute to this effect^{49–52}. Given the limited research on gender differences in FNP recovery, further studies are warranted.

This study limited the subjects to Bell's palsy, which usually has a favorable prognosis. Expanding the study to include Ramsay Hunt syndrome, which has a poorer prognosis, with a larger sample size may provide valuable insights into predicting recovery in intractable peripheral FNP associated with Ramsay Hunt syndrome.

The present study predicted future recovery and patient grades based on time-series data using SVM and LSTM. SVM is a widely used machine learning method based on the well-established mathematical concept of the margin of the classification hyperplane³⁴. Furthermore, LSTM, a type of RNN, is highly effective for handling sequential data^{35,36}. It was specifically designed to overcome the vanishing gradient problem, which traditional RNNs face during training. LSTM utilizes a cell state and a gate mechanism to selectively retain important information for a long period while discarding unnecessary information. This enables the model to learn both long-term and short-term dependencies within the data.

The limitations of this study include the small sample size and the limited dataset available for applying machine learning models. Although many patients with FNP visited the clinic, a significant number were lost to follow-up after favorable recovery, and some pursued alternative treatments after completing initial therapy. Consequently, the number of subjects—particularly in the validation set—was limited, restricting the evaluation of the classification model for patients with persistent FNP and potentially affecting the stability and generalizability of the results. In addition, patients with H–B grade 5 were not accurately predicted on SAD4 due to the low representation of such cases. Moreover, the dataset was imbalanced across classes because of the limited number of patients available for balanced sampling, and artificial balancing procedures, such as data synthesis or resampling, were not applied in order to preserve medical reliability. Despite this limitation, the present study demonstrated the potential of deep learning for predicting facial palsy. The distributions of the training and test datasets were similar, and the overall predictive performance of the model was satisfactory. Considering that the primary clinical objective of this study was the identification of recovered patients, the high recall for the recovery class indicates clinically acceptable discriminative performance, even with slight bias toward the majority class. As accuracy and mean absolute error improve with larger sample sizes, training and validating models on larger, more balanced datasets may provide better predictions for these patients. Further large-scale studies are warranted to confirm the clinical feasibility of this approach.

Conclusion

The present study demonstrated that changes in FNP grades can be accurately estimated, achieving an accuracy of 0.903 for predicting acceptable recovery and a MAE of approximately 0.46 for each clinic visit, using data from serial H–B grades from consecutive clinic visits and clinically relevant features. Given that FNP results in significant social and emotional costs, appropriate intervention is necessary based on the estimated prognosis. Our machine learning model may assist clinicians in predicting prognosis and providing appropriate consultation and treatment for FNP patients.

Data availability

The datasets analyzed during the current study are not publicly available due to the absence of IRB approval for sharing. However, they can be made available from the corresponding author upon reasonable request once IRB approval is granted.

Code availability

The proposed method was developed in Python (version 3.11), and the source code is publicly available at <https://github.com/parkjr0628/bell-palsy-ml>.

Received: 12 March 2025; Accepted: 31 December 2025

Published online: 06 January 2026

References

- Holland, N. J. & Weiner, G. M. Recent developments in Bell's palsy. *BMJ* **329**, 553–557. <https://doi.org/10.1136/bmj.329.7465.553> (2004).
- Mahadevappa, K., Vora, A., Graham, A. & Nesathurai, S. Facial paralysis: A critical review of accepted explanation. *Med. Hypotheses* **74**, 508–509. <https://doi.org/10.1016/j.mehy.2009.10.010> (2010).
- Jeong, J. et al. Monthly variation in Bell's palsy based on population data of Korea. *Audiol. Neurotol.* **29**, 290–296. <https://doi.org/10.1159/000536365> (2024).
- Finsterer, J. Management of peripheral facial nerve palsy. *Eur. Arch. Otorhinolaryngol.* **265**, 743–752. <https://doi.org/10.1007/s00405-008-0646-4> (2008).
- Yoo, M. C. et al. Evaluation of factors associated with favorable outcomes in adults with bell palsy. *JAMA Otolaryngol. Head Neck Surg.* **146**, 256–263. <https://doi.org/10.1001/jamaoto.2019.4312> (2020).
- Menchetti, I., McAllister, K., Walker, D. & Donnan, P. T. Surgical interventions for the early management of Bell's palsy. *Cochrane Database Syst. Rev.* **1**, Cd007468. <https://doi.org/10.1002/14651858.CD007468.pub4> (2021).
- Kim, Y. et al. Comparison of medical and surgical treatment in severe Bell's palsy. *J. Clin. Med.* **11**, 888 (2022).
- House, J. W. & Brackmann, D. E. Facial nerve grading system. *Otolaryngol. Head Neck Surg.* **93**, 146–147. <https://doi.org/10.1177/01945988509300202> (1985).
- Guntinas-Lichius, O. et al. Facial nerve electrodiagnostics for patients with facial palsy: A clinical practice guideline. *Eur. Arch. Otorhinolaryngol.* **277**, 1855–1874. <https://doi.org/10.1007/s00405-020-05949-1> (2020).
- Lee, D. H. Clinical efficacy of electroneurography in acute facial paralysis. *J. Audiol. Otol.* **20**, 8–12. <https://doi.org/10.7874/jao.2016.20.1.8> (2016).
- Lee, H.-J. et al. Correlation between House–Brackmann grade and electroneurography score in facial nerve paralysis patients. *Korean J. Audiol.* **12**, 105–110 (2008).
- Danielidis, V., Skevas, A., Van Cauwenberge, P. & Vinck, B. A comparative study of age and degree of facial nerve recovery in patients with Bell's palsy. *Eur. Arch. Otorhinolaryngol.* **256**, 520–522. <https://doi.org/10.1007/s004050050203> (1999).
- Tiemstra, J. D. & Khatkhate, N. Bell's palsy: Diagnosis and management. *Am. Fam. Physician* **76**, 997–1002 (2007).
- Seo, H. W. et al. Prediction of hearing recovery with deep learning algorithm in sudden sensorineural hearing loss. *Sci. Rep.* **14**, 20058. <https://doi.org/10.1038/s41598-024-70436-0> (2024).
- Park, J. R. et al. Data preprocessing and augmentation improved visual field prediction of recurrent neural network with multi-central datasets. *Ophthalmic Res.* **66**, 978–991. <https://doi.org/10.1159/000531144> (2023).
- Kim, H. et al. Visual field prediction using a deep bidirectional gated recurrent unit network model. *Sci. Rep.* **13**, 11154. <https://doi.org/10.1038/s41598-023-37360-1> (2023).
- Kina, E. TLEABLNN: brain and Alzheimer's disease detection using attention-based explainable deep learning and SMOTE using imbalanced brain MRI. *IEEE Access* **13**, 27670–27683. <https://doi.org/10.1109/ACCESS.2025.3539550> (2025).
- ten Harkel, T. C. et al. Optimization of the automated Sunnybrook facial grading system—improving the reliability of a deep learning network with facial landmarks. *Eur. Ann. Otorhinolaryngol. Head Neck Dis.* **142**, 5–10. <https://doi.org/10.1016/j.anorl.2024.07.005> (2025).

19. Kim, H. S., Kim, S. Y., Kim, Y. H. & Park, K. S. A smartphone-based automatic diagnosis system for facial nerve palsy. *Sensors* **15**, 26756–26768 (2015).
20. Nouis, A., O'Neil, R., Saif, A., Khatab, A. & Diallo, C. A hybrid Bi-LSTM model for data-driven maintenance planning. *Auton. Intell. Syst.* **5**, 13. <https://doi.org/10.1007/s43684-025-00099-9> (2025).
21. Xing, D., Wang, Y., Sun, P., Huang, H. & Lin, E. A CNN-LSTM-att hybrid model for classification and evaluation of growth status under drought and heat stress in chinese fir (*Cunninghamia lanceolata*). *Plant Methods* **19**, 66. <https://doi.org/10.1186/s13007-023-01044-8> (2023).
22. Doğan, E. LSTM training set analysis and clustering model development for short-term traffic flow prediction. *Neural Comput. Appl.* **33**, 11175–11188. <https://doi.org/10.1007/s00521-020-05564-5> (2021).
23. Boulmaiz, T., Guermoui, M. & Boutaghane, H. Impact of training data size on the LSTM performances for rainfall-runoff modeling. *Model. Earth Syst. Environ.* **6**, 2153–2164. <https://doi.org/10.1007/s40808-020-00830-w> (2020).
24. Kim, S.-Y. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinform.* **10**, 147. <https://doi.org/10.1186/1471-2105-10-147> (2009).
25. Zantvoort, K. et al. Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions. *NPJ Digit. Med.* **7**, 361. <https://doi.org/10.1038/s41746-024-01360-w> (2024).
26. Cerqueira, V., Torgo, L. & Mozetič, I. Evaluating time series forecasting models: an empirical study on performance estimation methods. *Mach. Learn.* **109**, 1997–2028. <https://doi.org/10.1007/s10994-020-05910-7> (2020).
27. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819. <https://doi.org/10.1136/bmj-2023-074819> (2024).
28. Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C. & van Diepen, M. External validation of prognostic models: What, why, how, when and where?. *Clin. Kidney J.* **14**, 49–58. <https://doi.org/10.1093/ckj/sfaa188> (2021).
29. Kudoh, A., Ebina, M., Kudo, H. & Matsuki, A. Delayed recovery of patients with Bell's palsy complicated by non-insulin-dependent diabetes mellitus and hypertension. *Eur. Arch. Otorhinolaryngol.* **255**, 166–167. <https://doi.org/10.1007/s004050050036> (1998).
30. Jeong, J., Yoon, S. R., Lim, H., Oh, J. & Choi, H. S. Risk factors for Bell's palsy based on the Korean national health insurance service national sample cohort data. *Sci. Rep.* **11**, 23387. <https://doi.org/10.1038/s41598-021-02816-9> (2021).
31. Psillas, G. et al. Evaluation of effects of diabetes mellitus, hypercholesterolemia and hypertension on Bell's palsy. *J. Clin. Med.* **10**, 2357. <https://doi.org/10.3390/jcm10112357> (2021).
32. Sullivan, F. M. et al. Early treatment with prednisolone or acyclovir in Bell's palsy. *N. Engl. J. Med.* **357**, 1598–1607. <https://doi.org/10.1056/NEJMoa072006> (2007).
33. Lagalla, G., Logullo, F., Di Bella, P., Provinciali, L. & Ceravolo, M. G. Influence of early high-dose steroid treatment on Bell's palsy evolution. *Neurol. Sci.* **23**, 107–112. <https://doi.org/10.1007/s100720200035> (2002).
34. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. <https://doi.org/10.1007/BF00994018> (1995).
35. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: continual prediction with LSTM. *Neural Comput.* **12**, 2451–2471. <https://doi.org/10.1162/089976600300015015> (2000).
36. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735> (1997).
37. Novak, C. B. Rehabilitation strategies for facial nerve injuries. *Semin. Plast. Surg.* **18**, 47–52. <https://doi.org/10.1055/s-2004-823123> (2004).
38. Nakano, H. et al. Physical therapy for peripheral facial palsy: A systematic review and meta-analysis. *Auris Nasus Larynx* **51**, 154–160. <https://doi.org/10.1016/j.anl.2023.04.007> (2024).
39. Goldstein, B. A., Pomann, G. M., Winkelmayer, W. C. & Pencina, M. J. A comparison of risk prediction methods using repeated observations: An application to electronic health records for hemodialysis. *Stat. Med.* **36**, 2750–2763. <https://doi.org/10.1002/sim.7308> (2017).
40. Painter, M. W. et al. Diminished schwann cell repair responses underlie age-associated impaired axonal regeneration. *Neuron* **83**, 331–343. <https://doi.org/10.1016/j.neuron.2014.06.016> (2014).
41. Chen, J. et al. Decreased blood vessel density and endothelial cell subset dynamics during ageing of the endocrine system. *Embo J.* **40**, e105242. <https://doi.org/10.15252/emboj.202105242> (2021).
42. Seals, D. R., Jablonski, K. L. & Donato, A. J. Aging and vascular endothelial function in humans. *Clin. Sci. (Lond)* **120**, 357–375. <https://doi.org/10.1042/cs20100476> (2011).
43. Li, X. et al. Inflammation and aging: Signaling pathways and intervention therapies. *Signal Transduct. Target. Ther.* **8**, 239. <https://doi.org/10.1038/s41392-023-01502-8> (2023).
44. Gonzalez-Freire, M., de Cabo, R., Studenski, S. A. & Ferrucci, L. The neuromuscular junction: Aging at the Crossroad between nerves and muscle. *Front Aging Neurosci.* **6**, 208. <https://doi.org/10.3389/fnagi.2014.00208> (2014).
45. Ikeda, M. et al. Clinical factors that influence the prognosis of facial nerve paralysis and the magnitudes of influence. *Laryngoscope* **115**, 855–860. <https://doi.org/10.1097/01.MLG.0000157694.57872.82> (2005).
46. Kasse, C. A. et al. The value of prognostic clinical data in Bell's palsy. *Braz. J. Otorhinolaryngol.* **71**, 454–458. [https://doi.org/10.1016/s1808-8694\(15\)31198-8](https://doi.org/10.1016/s1808-8694(15)31198-8) (2005).
47. Yon, D. K. et al. A Retrospective study on the gender differences in clinical manifestations of Bell's palsy. *Clin. Otolaryngol.* **50**, 307–315. <https://doi.org/10.1111/coa.14259> (2025).
48. Flifel, M. E., Belal, T. & Abou Elmaaty, A. A. Bell's palsy: Clinical and neurophysiologic predictors of recovery. *Egypt. J. Neurol. Psychiatry Neurosurg.* **56**, 40. <https://doi.org/10.1186/s41983-020-00171-6> (2020).
49. Blacklock, A. D., Cauveren, J. A. & Smith, P. G. Estrogen selectively increases sensory nociceptor innervation of arterioles in the female rat. *Brain Res.* **1018**, 55–65. <https://doi.org/10.1016/j.brainres.2004.05.075> (2004).
50. Gu, Y. et al. 17 β -estradiol enhances Schwann cell differentiation via the ER β -ERK1/2 signaling pathway and promotes remyelination in injured sciatic nerves. *Front Pharmacol.* **9**, 1026. <https://doi.org/10.3389/fphar.2018.01026> (2018).
51. He, B. et al. Factors predicting sensory and motor recovery after the repair of upper limb peripheral nerve injuries. *Neural Regen. Res.* **9**, 661 (2014).
52. Kovačić, U., Žele, T., Osredkar, J., Sketelj, J. & Bajrović, F. F. Sex-related differences in the regeneration of sensory axons and recovery of nociception after peripheral nerve crush in the rat. *Exp. Neurol.* **189**, 94–104. <https://doi.org/10.1016/j.expneurol.2004.05.015> (2004).

Acknowledgements

This study was supported by a clinical research grant funded by Seoul National University Boramae Medical Center (04-2023-0026). This study was also supported by a clinical research grant funded by Seoul National University Boramae Medical Center. This work was also supported by the research fund of Hanyang University (HY-202500000003190 and HY-202500000003704).

Author contributions

Sang-Yoon Han and Jeong Rye Park conducted statistical analyses, wrote the main manuscript, and prepared figures and tables. Heonjeong Oh and Sung-Min Park collected the data. Young Ho Kim and Jongyook Park

designed the project, supervised the project, and revised the manuscript. All authors reviewed the manuscript.

Funding

This study was supported by a clinical research grant funded by Seoul National University Boramae Medical Center (04-2023-0026). This study was also supported by a clinical research grant funded by Seoul National University Boramae Medical Center. This work was also supported by the research fund of Hanyang University (HY-202500000003190 and HY-202500000003704).

Competing interests

The authors declare no competing interests.

Ethics and informed consent statement

This study was approved by the Institutional Review Board of Seoul National University Boramae Medical Center (IRB No. 10-2022-90). As the present study was retrospective, the Institutional Review Board of Seoul National University Boramae Medical Center waived the requirement for informed consent.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-34934-z>.

Correspondence and requests for materials should be addressed to J.P. or Y.H.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026