Journal of Integrative Neuroscience

J. Integr. Neurosci. **2023**; 22(3): 57
https://doi.org/10.31083/j.jin2203057

Original Research

A Fully Automated Visual Grading System for White Matter Hyperintensities of T2-Fluid Attenuated Inversion Recovery Magnetic Resonance Imaging

ZunHyan Rieu[1,†], Regina EY Kim[1,†], Minho Lee[1], Hye Weon Kim[1], Donghyeon Kim[1], JeongHyun Yong[1], JiMin Kim[2], MinKyoung Lee[3], Hyunkook Lim[4], JeeYoung Kim[2,*]

[1]Research Institute, NEUROPHET Inc., 06234 Seoul, Republic of Korea
[2]Department of Radiology, Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 06247 Seoul, Republic of Korea
[3]Department of Radiology, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 06247 Seoul, Republic of Korea
[4]Department of Psychiatry, Yeouido St. Mary's Hospital, College of Medicine, The Catholic University of Korea, 06247 Seoul, Republic of Korea
*Correspondence: jeeyoungkim@catholic.ac.kr (JeeYoung Kim)
[†]These authors contributed equally.
Academic Editor: Gernot Riedel
Submitted: 30 September 2022 Revised: 21 December 2022 Accepted: 26 December 2022 Published: 6 May 2023

Abstract

Background: The Fazekas scale is one of the most commonly used visual grading systems for white matter hyperintensity (WMH) for brain disorders like dementia from T2-fluid attenuated inversion recovery magnetic resonance (MR) images (T2-FLAIRs). However, the visual grading of the Fazekas scale suffers from low-intra and inter-rater reliability and high labor-intensive work. Therefore, we developed a fully automated visual grading system using quantifiable measurements. **Methods**: Our approach involves four stages: (1) the deep learning-based segmentation of ventricles and WMH lesions, (2) the categorization into periventricular white matter hyperintensity (PWMH) and deep white matter hyperintensity (DWMH), (3) the WMH diameter measurement, and (4) automated scoring, following the quantifiable method modified for Fazekas grading. We compared the performances of our method and that of the modified Fazekas scale graded by three neuroradiologists for 404 subjects with T2-FLAIR utilized from a clinical site in Korea. **Results**: The Krippendorff's alpha across our method and raters (A) versus those only between the radiologists (R) were comparable, showing substantial (0.694 vs. 0.732; 0.658 vs. 0.671) and moderate (0.579 vs. 0.586) level of agreements for the modified Fazekas, the DWMH, and the PWMH scales, respectively. Also, the average of areas under the receiver operating characteristic curve between the radiologists (0.80 ± 0.09) and the radiologists against our approach (0.80 ± 0.03) was comparable. **Conclusions**: Our fully automated visual grading system for WMH demonstrated comparable performance to the radiologists, which we believe has the potential to assist the radiologist in clinical findings with unbiased and consistent scoring.

Keywords: Fazekas scale; white matter lesion hyperintensity; T2-FLAIR; deep-learning; brain segmentation

1. Introduction

T2-weighted fluid-attenuated inversion recovery magnetic resonance imaging (T2-FLAIRs) is used to assess the severity of white matter lesions that appeared as hyperintensities (WMHs) *in vivo*. WMH provides important information about brain health, aging, and possible disease burden [1–4]. WMH has been recognized as an important biomarker for small-vessel cerebrovascular diseases and Alzheimer's disease [5,6].

The Fazekas scale provides a conventional visual grading approach to quantify WMH severity into four scales and is often practiced by radiologists and in clinics worldwide [7]. The Fazekas scale classifies the severity of WMHs presented in the T2-FLAIR using the combination of the periventricular hyperintensity (PWMH) scale and the deep white matter hyperintensity (DWMH) scale [7]. Both PWMHs and DWMHs are graded from zero to three (Table 1) [7].

However, the use of the Fazekas scale in clinical practice or research is often limited by its labor-intensive process, as are all forms of visual grading [8], and low inter- and intra-rater reliability due to its ambiguous given criteria [9]. Over time, the age-related white matter changes (ARWMC) scale was introduced to overcome the ambiguousness of the subjectively measured Fazekas scale to provide quantifiable measurements [10]. Yet, the ARWMC scale also had limits due to not providing a detailed separation of DWMH and PWMH lesions. Hence, we had to find an advanced method that is computationally viable to implement for gratifying the original Fazekas scale. Several groups suggested a quantifiable method using the maximum diameter distance to divide DWMH and PWMH. The DWMH and PWMH scales are defined from the measured distance, which they call the modified Fazekas scale (Table 1) [11].

This study aims to provide an automated approach to the modified Fazekas scale that is efficient and easily applicable with reliable results in general clinical research and practice to assist doctors by reducing their labor-intensive

Table 1. Comparison of criteria between the original and the modified Fazekas scale.

	The original Fazekas scale	The modified Fazekas scale
Grade 0	Absent	Absent
Grade 1	PWMH: caps or pencil-thin lining DWMH: punctuate foci	PWMH $<$10 mm AND DWMH $<$10 mm
Grade 2	PWMH: smooth halo DWMH: beginning confluence	1. DWMH $<$10 mm AND PWMH $\geq$10 mm OR 2. 10 mm $\leq$ DWMH $<$25 OR 3. DWMH $\geq$25 mm AND PWMH $<$10 mm
Grade 3	PWMH: irregular periventricular signal extending into the deep white matter DWMH: large confluent areas	PWMH $\geq$10 mm AND DWMH $\geq$25 mm

DWMH, deep white matter hyperintensity; PWMH, periventricular hyperintensity.

process. Thus, this study shares our implementation and validation on a fully automated modified Fazekas scale using deep learning and a rule-based algorithm. Radiologists participated in this study to validate if our method is comparable to humans since this study is the first automation algorithm for the modified Fazekas scale.

2. Materials and Methods

2.1 Overview of the Proposed Method

The proposed approach consists of four stages (Fig. 1). First, the ventricle and WMH are segmented from the input 2D T2-FLAIR using a deep learning algorithm [12]. Second, the segmented WMHs are categorized into DWMHs and PWMHs following the rule suggested in the previous study [13]. Third, the maximum diameter is measured for both DWMHs and PWMHs according to the modified Fazekas scale. Finally, the modified Fazekas scale is calculated using the obtained maximum diameter of DWMH and PWMH. For validation, we compared the agreements of our proposed method against those of three certified radiologists.

2.2 Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Eunpyeon St. Mary's Hospital, College of Medicine, The Catholic University of Korea (IRB No. PC20EISI0094 on 02 July 2020).

2.3 Study Population Demographics

Two-dimensional (2D) T2-FLAIR scans from the Catholic University of Korea Eunpyeong St. Mary's Hospital were used in this study. The dataset was collected with the inclusion criteria of magnetic resonance imaging (MRI) containing WMH diagnosed with dementia. The exclusion criteria were WMHs with multiple pathologies, such as stroke or other disorders that may cause different components (e.g., cerebrospinal fluid, microbleeds) within the WMHs. The average age of the 404 participants was 68.7 $\pm$ 12.7 years.

2.4 MRI Acquisition

All images were acquired using a 3T MRI scanner (MAGNETOM Vida, Siemens Medical Solutions Inc., Malvern, PA, USA) with the following parameters: axial, time of echo (TE) = 114 ms, time of repetition (TR) = 8 s, time of inversion (TI) = 2370 ms, field of view (FOV) = 21 cm $\times$ 21 cm, slice thickness = 4 mm, number of slices = 32, with a gap = 1 mm, and acquisition matrix size = 384 $\times$ 230.

2.5 Comparison between Human Raters and Our Proposed Method

The modified Fazekas scale is based on measuring the maximum diameter (mm) of DWMH and PWMH, which is quantitative (Table 1). Theoretically, our computationally implemented measuring method would be more accurate than the human raters. Yet, we compared our automated results to the human raters to demonstrate the similarity since the main goal of developing this method is to help out the intense labor of humans. For human raters, each T2-FLAIR images were assessed by three certified radiologists with a subspecialty in neuroradiology. All patient information was blinded to make no bias in rating, and also that mutual information shall not be shared between the raters. The images were visually graded independently by raters following the criteria of the modified Fazekas scale. The raters manually used a MRI measuring tool to measure the diameter (mm) of the longest axis on the PWMH and DWMH. Measurement was done on raw MRI without any provided annotations. Then, radiologists provided the modified Fazekas scale on the basis of the measurement [11]. For our proposed method, we proceed with the automated pipeline shown in the overview of the proposed method (Fig. 1), then provide the modified Fazekas scale.

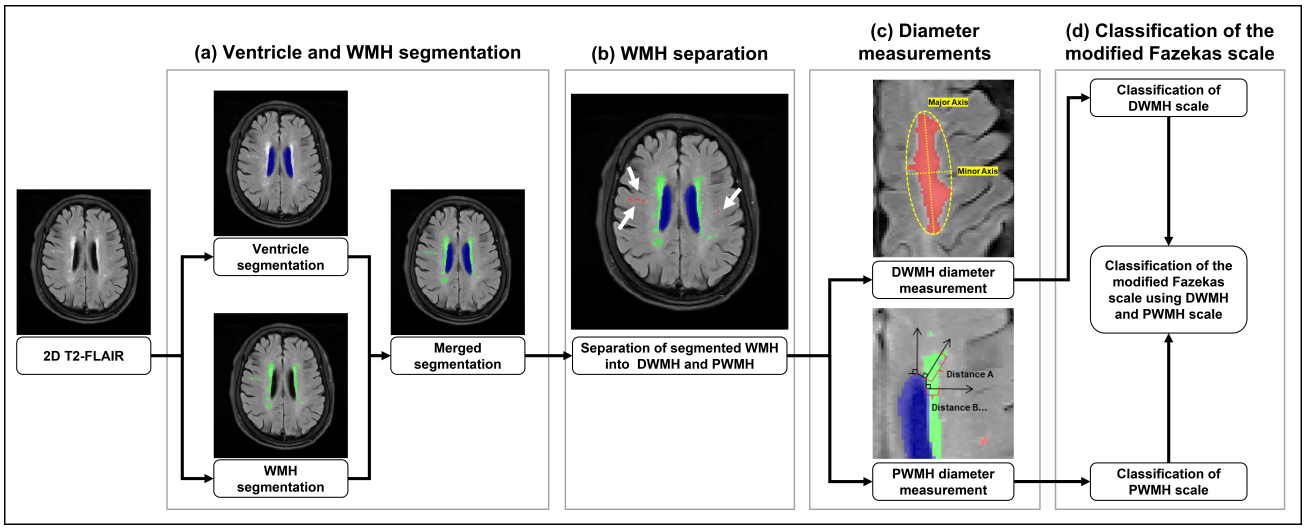

Fig. 1. The pipeline of the proposed method. Automated scoring for the Fazekas scale involves four stages and is based on T2-FLAIR MR images. (a) Brain tissue and WMH segmentation. (b) WMH separation. (c) Diameter measurement. (d) Fazekas scale prediction.

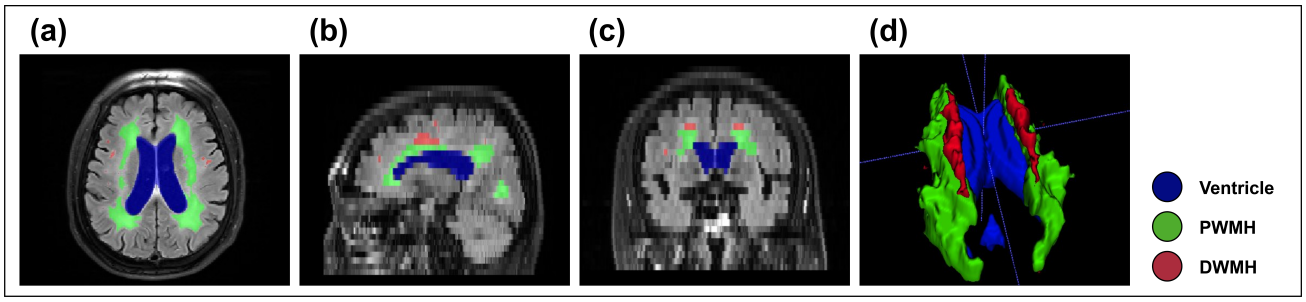

Fig. 2. DWMH and PWMH separation results in multiple planes. The blue, green, and red labels represent the ventricles, PWMH, and DWMH segmentations, respectively. (a) Axial plane. (b) Sagittal plane. (c) Coronal plane. (d) 3D view from the top.

2.6 Automated Classification for the Modified Fazekas Scale

2.6.1 T2-FLAIR Segmentation between Ventricle and WMH

We used our previously reported in-house method for simultaneous ventricle and WMH segmentation (Fig. 1a) [12]. The publication introduced two individual deep learning-based segmentation methods for T2-FLAIR. This research aimed to produce brain tissues and WMH segmentation using T2-FLAIR without its paired T1-weighted MRI (T1). We utilized the semi-supervised learning method and constructed the deep learning-based segmentation model to train FreeSurfer-generated brain tissue, including the ventricle from T1 to T2-FLAIR [14,15]. Then, the WMH model was trained with U-Net-based architecture using manually annotated and clinically confirmed WMH labels from radiologists utilizing PyTorch (version 1.7.1, python software foundation, Wilmington, DE, USA) [16, 17]. The previous research datasets are unrelated to our automated approach. The in-house segmentations demonstrated promising results for further clinical relevance and application.

All processed segmentation labels from the models used for this study were set to right-anterior-superior (RAS) orientation and resampled to 1×1 mm^3 spacing for the axial plane. Then, the ventricle and WMH segmentation results were merged for further measurement.

2.6.2 WMH Separation into DWMH and PWMH

We categorized the segmented WMH region further into DWMH and PWMH regions (Fig. 2). The separation was based on the calculated distance between the DWMHs/PWMHs and the boundaries of the segmented ventricle regions. For the X and Y axes, we separated PWMHs and DWMHs in 2D slice-based where ventricle segmentation exists in the axial plane: PWMHs were specified from WMHs within $\leq$13 mm from the margin of the ventricles; DWMHs were specified from WMHs outside of $>$13 mm [13]. For the Z-axis, we defined PWMHs and DWMHs based on the range of the ventricles in the Z-axis: PWMHs for WMHs from the lowest slice to slice one above the ventricle and DWMHs for others [11].

2.6.3 Diameter of DWMH and PWMH

We measured the diameters of the separated DWMH and PWMH (Fig. 3). The vertical distance was used for DWMHs, and the horizontal distance was used for PWMHs, as suggested in the modified Fazekas scale [11].

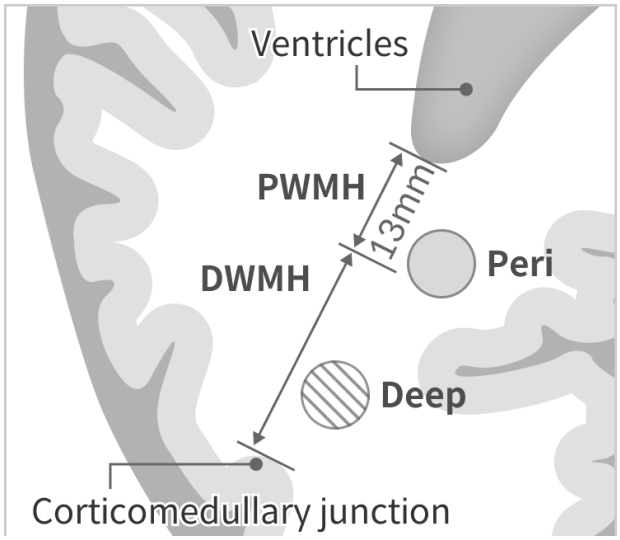

Fig. 3. Overview of the WMH separation into DWMH and PWMH. WMH, white matter hyperintensity; DWMH, deep white matter hyperintensity; PWMH, periventricular white matter hyperintensity.

Principal Component Analysis (PCA) based on the euclidean distance was performed on DWMHs in all 2D axial planes to measure the vertical diameter [18]. Taking the irregularly shaped DWMH as an input, the PCA-based measurement generates an approximated ellipse around the DWMH (Fig. 4d). Then, the major and minor axes are suggested for the eclipse. Since the DWMH scale is measured from the maximum diameter, we utilized the distance of the major axis [18].

PWMH is measured by measuring the horizontal diameter between the ventricle and the PWMH. Since the horizontal diameter varies from the starting point of the ventricle, we created a 2D Danielsson distance map for all 2D axial slices containing PWMHs and ventricles (Fig. 5) [19]. We extracted the ventricle contour from the distance map. We created perpendicular rays with a length of 13 mm from each pixel coordinate of the ventricle contour, representing the cut-off distance between PWMH and DWMH [13]. For each cluster of PWMH, we measured the mean distance of every ray that intersected the PWMH.

2.6.4 Classification of the Modified Fazekas Scale

At this stage (Fig. 1d), we finalized the automation process by classifying the modified Fazekas scale. Using the measured maximum diameters of the DWMHs and PWMHs, we assigned scales ranging from 1 to 3 (Table 1) as suggested by the modified Fazekas scale [11]. For the PWMHs, 1 represented maximum diameters <5 mm, 3 represented maximum diameters ≥ 10 mm, and 2 represented maximum diameters ≥ 5 mm and <10 mm. For the DWMHs, 1 represented maximum diameters <10 mm, 3 represented maximum diameters ≥ 25 mm, and 2 represented maximum diameters ≥ 10 mm but <25 mm. Finally, we classified the modified Fazekas scale using the WMH Visual rating system (Table 1).

2.7 Performance Evaluation

We investigated the agreements of the modified Fazekas scale from our proposed method and the experts with different years of experience. The multiple-rater agreement was assessed using Krippendorff's alpha [20]. Krippendorff's alpha was utilized to provide the level of agreement between the visual gradings performed by the radiologists and our proposed method. The inter-rater agreement was assessed using the areas under the receiver operating characteristic curves (AUROCs) [21] for the proposed method and each radiologist assessment. The AUROC was utilized to present the correspondence between our proposed method and the radiologists. The AUROC was used to determine the decision threshold for the classification performance of the two raters related to the true-positive rate (TPR) and false-positive rate (FPR) within the range of 0 to 1. Higher AUROCs are associated with higher performance than the gold standard [21]. All the performance evaluation was conducted either using R package software version 3.4.3 (The R Foundation for Statistical Computing, Vienna, Austria) or Python version 3.7 (Python Software Foundation) with the scikit-learn library [22–24].

3. Results

3.1 Multiple-Rater Agreement

To investigate the level of agreement between the different ratings, we assessed the multiple-rater agreement using Krippendorff's alpha (α) [25]. The multiple-rater agreements (α) with and without our proposed method for the DWMH scale, PWMH scale, and the modified Fazekas scale are shown in Table 2. The agreement of the modified Fazekas scale among the radiologists (R) and the ratings including our proposed method (A) were both substantial, as indicated by $\alpha = 0.732$ and 0.694, respectively, as suggested in Krippendorff's alpha [20]. The multi-rater agreement (α) was also substantial (R, 0.671; A, 0.658) for DWHH and moderate (R, 0.586; A, 0.579) for PWMH, as suggested in Krippendorff's alpha [20]. Note that the multi-rater agreement (α) for among the radiologists' ratings only (R) was consistently higher (the modified Fazekas scale, +0.038; DWMH scale, 0.013; PWMH scale, 0.007) than the agreement of the radiologists' ratings and our proposed method.

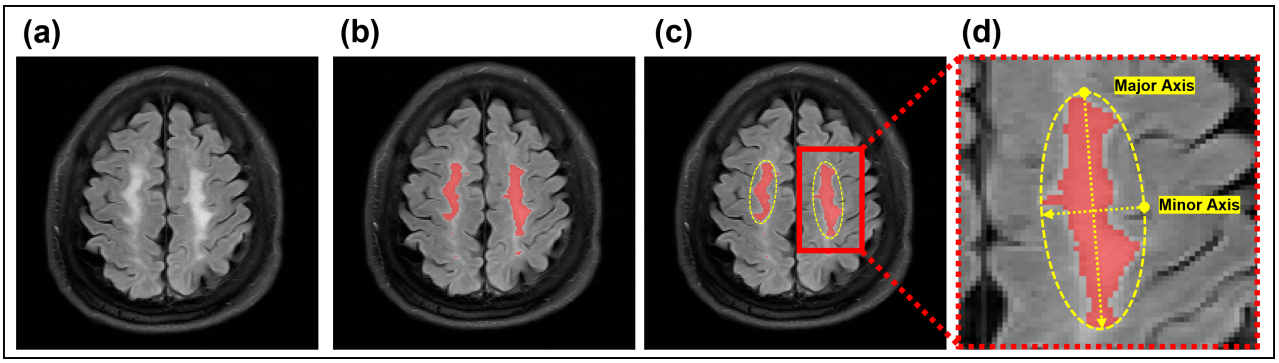

Fig. 4. Measurement of DWMH performed with PCA method. (a) T2-FLAIR MRI input. (b) DWMH segmentation. (c) PCA method. (d) Calculation of the major and minor axes.

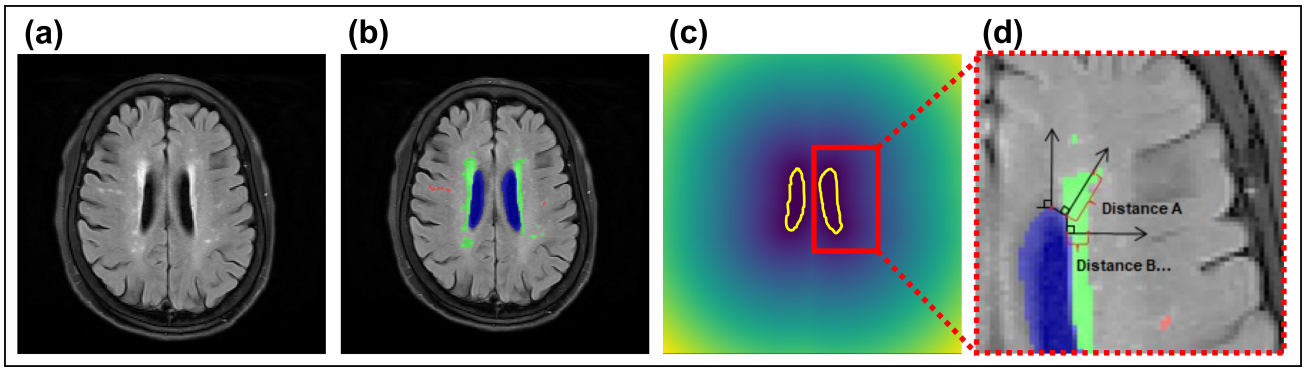

Fig. 5. Measurement of PWMH with four stages. (a) T2-FLAIR MRI input. (b) Combined segmentation results with ventricles and WMHs. (c) Distance map from ventricle segmentation. (d) PWMH measurement using ventricle segmentation and a distance map.

Table 2. Multiple-rater agreement using Krippendorff's alpha coefficient for the modified Fazekas scale.

ROI	Multiple-rater agreements (α)	
	(R) without proposed method R1, R2, and R3	(A) with proposed method R1, R2, R3, and P
The modified Fazekas scale	0.732*	0.694*
DWMH scale	0.671*	0.658
PWMH scale	0.586	0.579

ROI, region of interest; DWMH, deep white matter hyperintensity; PWMH, periventricular hyperintensity; P, proposed method; R1/2/3, raters 1, 2, and 3; *, Krippendorff's alpha (α) greater than ($\geq$) 0.667.

3.2 Inter-Rater Agreement

We determined the performance agreement using AUROCs. The agreements of the modified Fazekas scales determined by the radiologists and the proposed method are summarized in Table 3: G shows the evaluations by the radiologists (R1 vs. R2, R1 vs. R3, and R2 vs. R3), and M shows the evaluations by the raters and the proposed method (R1 vs. P, R2 vs. P, R3 vs. P). The interpretations of the area under the curve (AUROC) coefficients are as follows: 0.5, no discrimination; 0.6 to 0.7, poor discrimination; 0.7 to 0.8, acceptable discrimination; 0.8 to 0.9, excellent discrimination; 0.9 to 1.0, outstanding discrimination [26]. The average AUROC scores for the modified Fazekas scale determined by the radiologists showed excellent discrimination (G 0.87 $\pm$ 0.06; M 0.83 $\pm$ 0.05) for the modified Fazekas scale 1, excellent and acceptable discrimination (G 0.83 $\pm$ 0.08; M 0.77 $\pm$ 0.05) for the modified Fazekas scale 2, and acceptable discrimination (G 0.70 $\pm$ 0.10; M 0.79 $\pm$ 0.09) for the modified Fazekas scale 3. The average AUROC score for the agreement between the radiologists (G) was higher than that for our proposed method (M), the modified Fazekas scale 1 (+0.04), and the modified Fazekas scale 2 (+0.06). In contrast, M showed a higher score than G for the modified Fazekas scale 3 (+0.09).

4. Discussion

In this study, we demonstrated a fully automated visual grading system for WMH using the modified Fazekas

Table 3. Inter-rater agreement using AUROC coefficient on the modified Fazekas scale.

(G) between radiologists				(M) against our proposed method			
	Modified Fazekas scale				Modified Fazekas scale		
	1	2	3		1	2	3
R1 vs. R2	0.89	0.85	0.63	R1 vs. P	0.81	0.74	0.71
R1 vs. R3	0.80	0.74	0.81	R2 vs. P	0.79	0.75	0.89
R2 vs. R3	0.93	0.91	0.64	R3 vs. P	0.88	0.83	0.78
Average	0.87 ± 0.06	0.83 ± 0.08	0.70 ± 0.10	Average	0.83 ± 0.05	0.77 ± 0.05	0.79 ± 0.09
Combined		0.80 ± 0.09		Combined		0.80 ± 0.03	

The inter-rater agreements of raters only (G, left part of the table) and raters vs. proposed method (M, right part of the table); R1/2/3, raters 1, 2, and 3; P, proposed method.

scale on T2-FLAIRs. Our approach aimed to automate the visual grading of the modified Fazekas scale utilizing deep learning and rule-based algorithms with quantifiable imaging-driven measurements using T2-FLAIR exclusively. This study was the first attempt to automate the WMH visual grading using the modified Fazekas scale [11].

Theoretically, since our proposed method is a computational implementation, it is more accurate than the manually calculated results from the human raters when it comes to measuring the diameter of WMHs. Nevertheless, performance evaluations were done on comparing our results to the radiologists' assessments, mainly due to two big reasons. First, the main goal of this method is to help doctors on reducing labor time and cost on daily basis. Second, since we are the first software to implement the modified Fazekas scale, comparison with other software was impossible. Hence, we compared our proposed method to human raters with multiple-rater and inter-rater agreements, which showed a high correspondence. Further investigation of the intra correlation coefficient (ICC) between software is preferred [27].

The multiple-rater agreement investigation (rating agreements with and without our proposed method suggested that the level of agreement from our approach was comparable to those among the radiologists. We used Krippendorff's alpha (α), which indicates the reliabilities of multiple raters for multiple categories [28]. Our results indicated an agreement between the radiologists was similar to the agreement between the radiologists and our proposed method (Table 2). We noticed '(A) with the proposed method R1, R2, R3, and P' had slightly lower agreement than '(R) without proposed method R1, R2, and R3' for all scales. The lower agreement with our proposed tool is due to the nature of Krippendorff's alpha, as the formula contains the weights on the number of raters in the denominator [20].

The inter-rater agreement between the radiologists and our proposed method demonstrated an equivalent performance on AUROC as well, which indicates the classification performance of the modified Fazekas scale between the two raters. The average AUROC showed minimal differences in the comparisons within radiologists (G) and between the radiologists and our proposed method (M) for the modified Fazekas scale 1 (G 0.87 vs. M 0.83), the modified Fazekas scale 2 (G 0.83 vs. M 0.77), and the modified Fazekas scale 3 (G 0.70 vs. M 0.79).

The average AUROC coefficient being higher in lower modified Fazekas scale means that the radiologists performed better for small WMH burdens than our proposed method. In contrast, our proposed method performed better than all of each radiologist and also the average AUROC coefficient for grade 3 for the modified Fazekas scale. This indicates out method may be clinically useful for objective disease severity evaluation in large WMH burdens. Regardless, the combined AUROC of the modified Fazekas scales demonstrated that the performance value between G and P was comparable (G 0.80 ± 0.09 vs. P 0.80 ± 0.03), suggesting that our proposed method is clinically useful as an objective indicator for WMH evaluation.

Our study has a few limitations. The implemented modified Fazekas scale may not be widely used more than the original version. However, since the original Fazekas scale is not quantifiable and is based on a qualitative and subjective grading, we had to implement a scale which is applicable to automatic analysis. Additionally, our proposed system is currently being developed, and it has been mostly tested using 2D T2-FLAIRs. While this approach can be extended to any T2-FLAIR protocol, its performance may vary depending on the protocol. Future validation studies are needed to generalize our approach. Another limitation is the lack of ground-truth data, which is grand-scale collected data, on the modified Fazekas scale. We validated our approach against the three radiologists, whose results were used as the standard for comparison. As we have observed from our results, the three radiologists did not agree perfectly, and the ground-truth for the modified Fazekas scale has not been established at this point. To overcome the lack of ground-truth, further studies involving more experienced experts are needed to establish the gold standard for the modified Fazekas scale.

This study presented an automated modified Fazekas scoring approach using the objective measurements driven from T2-FLAIR and showed its performance against certified neuroradiologists. More work is needed to show our

approach's applicability to the research and clinical setting in the near future. Even so, we believe the present work could also contribute to both scientific society and clinical environments by suggesting automated analysis for the modified Fazekas scoring, especially for research related to large-scale or multi-site of WMH.

5. Conclusions

We introduced a fully automated visual grading system for WMH of T2-FLAIRs based on deep learning and rule-based algorithms utilizing the modified Fazekas scale. As we aimed, the results of our method were comparable to those of the three certified radiologists who used the visual grading method. We believe that our proposed method may assist clinic works and radiologists' reading with its fully automated and quantifiable Fazekas scale with consistent measurement.

Availability of Data and Materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author Contributions

ZR, REK, ML, DK, and JYK designed the research study. ZR, REK, ML, JYK, JMK, MKL, HL, and JY performed the research. ZR, REK, and HWK contributed to the interpretation of the results. ZR, REK, and JYK analyzed the data. ZR, REK, and JYK wrote the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work and agreed to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of Eunpyeong St. Mary's Hospital, College of Medicine, The Catholic University of Korea (IRB No. PC20EISI0094 on 02 July 2020). The authors confirm that all subjects or legally authorized representatives signed written informed consent forms.

Acknowledgment

Not applicable.

Funding

This work was supported by the Korea Medical Device Development Fund grant funded by the Korea government (Project Number: 202015X34, KMDF-PR-20200901-0306).

Conflict of Interest

The authors declare no conflict of interest. Zun-Hyan Rieu, Regina EY Kim, Minho Lee, Hye Weon Kim, Donghyeon Kim and JeongHyun Yong belong to Research Institute, NEUROPHET Inc. The authors declare that there have no conflicts of interest.

References

[1] Garnier-Crussard A, Bougacha S, Wirth M, André C, Delarue M, Landeau B, *et al*. White matter hyperintensities across the adult lifespan: relation to age, Aβ load, and cognition. Alzheimer's Research & Therapy. 2020; 12: 127.

[2] Capizzano AA, Ación L, Bekinschtein T, Furman M, Gomila H, Martínez A, *et al*. White matter hyperintensities are significantly associated with cortical atrophy in Alzheimer's disease. Journal of Neurology, Neurosurgery, and Psychiatry. 2004; 75: 822–827.

[3] Enzinger C, Fazekas F, Matthews PM, Ropele S, Schmidt H, Smith S, *et al*. Risk factors for progression of brain atrophy in aging: six-year follow-up of normal subjects. Neurology. 2005; 64: 1704–1711.

[4] Brickman AM, Honig LS, Scarmeas N, Tatarina O, Sanders L, Albert MS, *et al*. Measuring cerebral atrophy and white matter hyperintensity burden to predict the rate of cognitive decline in Alzheimer disease. Archives of Neurology. 2008; 65: 1202–1208.

[5] Ferreira D, Shams S, Cavallin L, Viitanen M, Martola J, Granberg T, *et al*. The contribution of small vessel disease to subtypes of Alzheimer's disease: a study on cerebrospinal fluid and imaging biomarkers. Neurobiology of Aging. 2018; 70: 18–29.

[6] Guerrero R, Qin C, Oktay O, Bowles C, Chen L, Joules R, *et al*. White matter hyperintensity and stroke lesion segmentation and differentiation using convolutional neural networks. NeuroImage: Clinical. 2018; 17: 918–934.

[7] Fazekas F, Chawluk JB, Alavi A, Hurtig HI, Zimmerman RA. MR signal abnormalities at 1.5 T in Alzheimer's dementia and normal aging. American Journal of Neuroradiology. 1987; 8: 421–426.

[8] Manouvelou S, Koutoulidis V, Tolia M, Gouliamos A, Anyfantakis G, Moulopoulos LA, *et al*. Differential diagnosis of Alzheimer's disease and vascular dementia using visual rating scales. Hellenic Journal of Radiology. 2020; 5: 2–9.

[9] Valdés Hernández MDC, Morris Z, Dickie DA, Royle NA, Muñoz Maniega S, Aribisala BS, *et al*. Close correlation between quantitative and qualitative assessments of white matter lesions. Neuroepidemiology. 2013; 40: 13–22.

[10] Wahlund LO, Barkhof F, Fazekas F, Bronge L, Augustin M, Sjögren M, *et al*. A new rating scale for age-related white matter changes applicable to MRI and CT. Stroke. 2001; 32: 1318–1322.

[11] Moon SY, Na DL, Seo SW, Lee J, Ku BD, Kim SY, *et al*. Impact of white matter changes on activities of daily living in mild to moderate dementia. European Neurology. 2011; 65: 223–230.

[12] Rieu Z, Kim J, Kim RE, Lee M, Lee MK, Oh SW, *et al*. Semi-Supervised Learning in Medical MRI Segmentation: Brain Tissue with White Matter Hyperintensity Segmentation Using FLAIR MRI. Brain Sciences. 2021; 11: 720.

[13] Kim KW, MacFall JR, Payne ME. Classification of white matter lesions on magnetic resonance imaging in elderly persons. Biological Psychiatry. 2008; 64: 273–280.

[14] Zhu X, Goldberg AB. Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning 3.1. 2009; 3: 1–130.

[15] Fischl B. FreeSurfer. NeuroImage. 2012; 62: 774–781.
[16] Olaf R, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Springer: Cham. 2015.
[17] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, *et al*. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32. NeurIPS. 2019.
[18] Elmore KL, Richman MB. Euclidean distance as a similarity metric for principal component analysis. Monthly Weather Review. 2001; 129: 540–549.
[19] Danielsson PE. Euclidean distance mapping. Computer Graphics and Image Processing. 1980; 14: 227–248.
[20] Krippendorff K. Content Analysis: An Introduction to Its Methodology. Sage publications: USA. 1980.
[21] Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. Indian Pediatrics. 2011; 48: 277–287.
[22] Team, R. Core. R: A language and environment for statistical computing. 2013. Available at: http://www.R-project.org/ (Accessed: 31 December 2022).
[23] Van Rossum G, Drake FL. Python 3 reference manual. CreateSpace: Scotts Valley, CA. 2009.
[24] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–2830.
[25] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977; 33: 159–174.
[26] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. John Wiley & Sons: USA. 2013.
[27] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine. 2016; 15: 155–163.
[28] Zapf A, Castell S, Morawietz L, Karch A. Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology. 2016; 16: 93.