

Automated Label Detection and Recommendation System Using Deep Convolution Neural Networks Using ANOVA

Sudhakara Reddy Peram*

Engineering Leader, Illumio Inc., United State

Abstract

An automated label identification and recommendation system using deep convolutional neural networks, understanding user perceptions regarding the reliability, accuracy, and interpretability of the computer system is essential for developing robust and explainable AI solutions. This study contributes to this need by exploring several performance dimensions, including accuracy, contextual relevance, clarity, dataset quality, sensitivity, specificity, recall rate, false positive rate, and explainability. Using a cross-sectional survey design, data were collected from respondents via Google Forms in November-December 2025. One-way ANOVA and two-way ANOVA ($p < 0.05$) were used with IBM SPSS to explore the relationships between socio-demographic variables and perceived performance metrics. The findings indicate consistent perceptions across experience levels and model types, while the model output format significantly impacted label accuracy and dataset quality. The results emphasize the importance of output representation in improving automated labeling performance and reliability, suggesting directions for future system optimization and validation frameworks.

Key Words: Automatic labeling, deep learning, CNN, model output, ANOVA, interpretability

Introduction

Automated labeling systems powered by deep convolutional neural networks (CNNs) have emerged as transformative solutions in diverse fields, from medical imaging to industrial applications [1]. These systems address a fundamental challenge in machine learning: the time-consuming, expensive, and often subjective nature of manual data annotation. By leveraging the pattern recognition capabilities of CNNs, automated labeling systems can process vast amounts of data with a consistency and speed unmatched by human annotators, while simultaneously reducing costs and enabling scalability for large-scale applications [2]. These networks learn increasingly complex features in deeper layers. Modern approaches often utilize pre-trained models like ResNet or DenseNet as backbone networks, which have demonstrated superior performance in feature extraction tasks [3]. These networks process input images through multiple convolutional and pooling layers, automatically learning spatial hierarchies of features without manual engineering. The architecture often includes mechanisms that focus computational resources on relevant image regions, improving both accuracy and interpretability [5]. Advanced systems integrate multiple detection

stages, combining regional proposal networks with classification heads to simultaneously detect and classify objects. This two-stage approach, exemplified by Faster R-CNN architectures, has proven particularly useful for scenarios requiring accurate localization along with classification [6]. The incorporation of batch normalization, dropout layers, and residual connections helps prevent overfitting while enabling the training of very deep networks that capture the subtle discriminative features necessary for accurate labeling.

The success of automated labeling systems depends on data quality and preprocessing pipelines. Research has shown that even small percentages of incorrectly labeled training data can significantly degrade model performance [7]. The challenge of label noise has been extensively documented in medical imaging applications. [8] Di Noto et al. (2020) demonstrated that automated label cleaning systems achieved 95% accuracy in pneumonia detection through pediatric chest X-ray analysis, surpassing previously reported models by 92% [9]. Their unsupervised data cleaning (UDC) technique successfully identified both incorrect labels and noisy data, where neither AI nor radiologists could consistently classify the images, highlighting the dual challenge of mislabeling and inherent data ambiguity. Similarly, Lin et al. (2025) addressed label noise in retinal image datasets, with their CleanLab-based approach improving label accuracies from a baseline of 62.9% after six cleaning cycles, accurately correcting most label errors (86.6–97.5%) [10]. Researchers have found that deep convolutional neural networks perform exceptionally well in automated detection applications. In 2018, Fang and his colleagues developed an improved version of Faster R-CNN specifically for monitoring construction sites. When used in real-world environments, it achieved an accuracy of 91% in detecting workers and 95% in detecting excavators. [11]. Their research showed

Received date: December 05, 2025 **Accepted date:** December 09 2025; **Published date:** December 15, 2025

***Corresponding Author:** Peram, S. R., Engineering Leader, Illumio Inc., United State; E-mail: sudhakarap2013@gmail.com

Copyright: © 2025 Peram, S. R. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Citation: Peram, S. R. (2025). Automated Label Detection and Recommendation System Using Deep Convolution Neural Networks Using ANOVA. International Journal of Artificial Intelligence and Machine Learning, 3(4), 1-8. <https://doi.org/10.55124/jaim.v3i4.294>

that combining spatial degradation assessments with automated classification systems significantly improves diagnostic capabilities compared to traditional manual feature extraction techniques [12].

The study emphasized that even a fraction of substandard data can significantly hinder AI performance, thus necessitating robust data quality assessment algorithms. The integration of multiple data modalities in automated labeling systems has emerged as a significant advancement [13]. Cui and colleagues (2021) integrated structural brain volume measurements from T1-weighted MRI scans with DTI-based metrics to identify amnestic mild cognitive impairment, yielding 71.09% accuracy, 51.96% sensitivity, and 78.40% specificity [14]. Their research, using the FreeSurfer-initiated large deformation diffeomorphic metric mapping (FS+LDDMM) technique, showed that combining structural brain morphological changes with white matter alterations improves the detection of subtle brain abnormalities compared to single-modality imaging approaches.[15]. This study found that various socio-demographic, lifestyle, and health variables influence classification disorders. This highlights the need for comprehensive feature selection strategies.[16]. Despite significant advances in automated labeling systems, current approaches face critical limitations [17-18]. Existing methods predominantly operate in isolation, lacking integration of ensemble detection with multi-modal data fusion and adaptive weak supervision mechanisms [19]. Furthermore, there is insufficient attention to real-time recommendation systems that not only detect labels but also suggest optimal labeling strategies based on data characteristics, quality assessment, and domain-specific requirements. [20] The absence of comprehensive frameworks that simultaneously address label noise detection, quality assessment, and intelligent label recommendation while maintaining interpretability represents a significant gap in current research. The objective of this study is to explore the impact of deep learning expertise on the descriptive capabilities of automated labeling systems.

Methodology

This research was conducted during November-December 2025 using a structured online questionnaire administered via Google Forms. Independent variables included gender, occupation, experience with deep learning models, application domain, deep learning model type, and model output format (2D and 3D representations) (Table 1). Dependent variables comprised perceived accuracy, contextual relevance, clarity, dataset quality, uniqueness, sensitivity, false positive rate, recall, and interpretability of a label. Data were collected using a five-point Likert scale, and the researchers used the ANOVA method along with other statistical methods for the analysis. The dependent variables reflected the respondents' opinions on various dimensions of the deep learning system's performance. These included label accuracy, contextual relevance, clarity, dataset quality, uniqueness, sensitivity, false positive rate, recall, and interpretability. Attitudinal variables were measured using a five-point Likert scale ranging from very poor to very good; this recorded users' assessments of model performance, reliability, and comprehensibility. To maximize participant reach, the Google Forms survey link was distributed through digital channels including Facebook, WhatsApp, Instagram, and LinkedIn. Respondents were informed about the survey objectives, and participation was voluntary, anonymous, and with assured confidentiality.

The online method facilitated convenient and cost-effective data collection from diverse professional and demographic groups. The automated label detection performance metric showed acceptable internal consistency. One-way ANOVA results did not reveal statistically significant differences in the perceived performance metrics across different levels of deep learning experience or various deep learning model types (Tables 3 and 5), indicating consistent perceptions regardless of expertise or architecture. In contrast, the analysis based on the model output format identified statistically significant differences in label accuracy and dataset quality between 2D and 3D patch outputs (Table 6), underscoring the importance of output representation. Two-way ANOVA further confirmed the absence of significant main or interaction effects (Tables 7). All analyses were conducted using IBM SPSS 27.0 software.

Table 1. Independent Variables and Their Response Options Used in the Study

Gender	Male/Female/Others
Occupation	Industry/Academics/Scientist/R&D person/others
Experience In Deep Learning Model	below a year/1-2 year/2-4 year/4-6 year/6 year above
Domain	chemical/medical/civil/mechanical/textile/pharmacy/others
DL model	ResNet/3D ResNet/CNN/RCNN/nnDetection/DeepMedic/3D UNET/ResNet/HeadXNet/2D CNN/1DCNN/SPCNet/Unet/DeepLabv3+/WLBiLSTM/CLCNN/CL+WLBiLSTM
Model Output	2D patches/3D patches

Table 2. Socio-demographic Characteristics of Respondents

Characteristics	Frequency	Percent (%)
Gender		
Male	191	35.6
Female	279	52
Others	67	12.5
Occupation		

Industry	98	18.2
Academics	158	29.4
Scientist	157	29.2
R&D person	101	18.8
others	23	4.3
Experience In Deep Learning Model		
below a year	36	6.7

1-2 years	125	23.3
2-4 years	199	37.1
4-6 years	144	26.8
6 years above	33	6.1
Domain		
chemical	56	10.4
medical	93	17.3
civil	75	14
mechanical	150	27.9
textile	71	13.2
pharmacy	78	14.5
others	14	2.6
DL model		
ResNet	14	2.6
3D ResNet	19	3.5
CNN	30	5.6

RCNN	56	10.4
nnDetection	42	7.8
DeepMedic	37	6.9
3D UNET	29	5.4
HeadXNet	38	7.1
2D CNN	41	7.6
1D CNN	32	6
SPCNet	46	8.6
UNet	29	5.4
DeepLabv3+	54	10.1
WLBiLSTM	31	5.8
CLCNN	24	4.5
CL+WLBiLSTM	15	2.8
Model Output		
2D patches	343	63.9
3D patches	194	36.1

Table 2 shows the socio-demographic characteristics of the survey participants; the majority of them are female participants working in the fields of education and science. The majority of them had 2–4 years of deep learning experience and worked in the mechanical and medical fields. CNN-based architectures were commonly used, and 2D patch outputs were preferred over 3D patches.

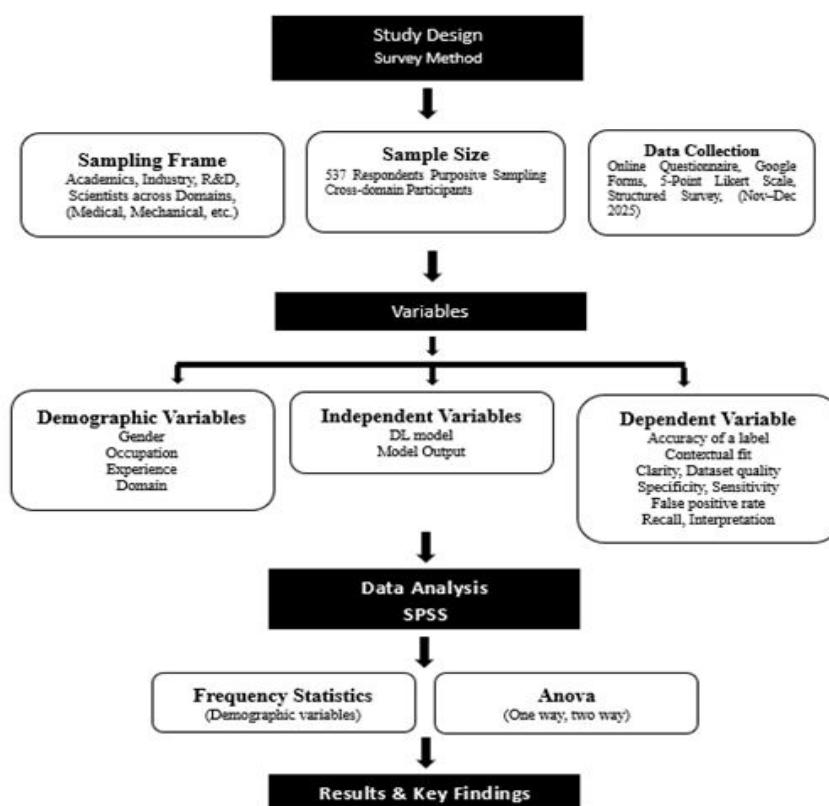


Figure 1: Research Methodology Flowchart

Table 3. One-Way ANOVA Examining Differences in Performance Metrics Across Experience Levels in Deep Learning Models

	Sum of Squares	df	Mean Square	F	Sig.
Accuracy of a Label	3.684	4	0.921	0.850	0.494
Contextual Fit	2.299	4	0.575	0.413	0.799
Clarity	8.016	4	2.004	1.198	0.311
Dataset Quality	2.069	4	0.517	0.314	0.869
specificity	9.715	4	2.429	1.290	0.273
Sensitivity	1.235	4	0.309	0.180	0.949
False Positive Rate	2.391	4	0.598	0.404	0.806
Recall	3.367	4	0.842	0.534	0.711
Interpretation	8.294	4	2.074	1.167	0.324

Table 3 presents the results of an ANOVA, a method used to examine differences in various performance metrics across experience levels in deep learning models. Since all p-values exceeded the 0.05 significance level, these results indicate that there were no statistically significant differences in any of the measured variables between the experimental groups. Metrics such as label accuracy, contextual relevance, clarity, dataset quality, uniqueness, sensitivity, false positive rate, recall, and interpretability remained relatively consistent regardless of experience. These findings suggest that the level of experience in deep learning did not have a significant influence on the performance outcomes observed or measured in this study.

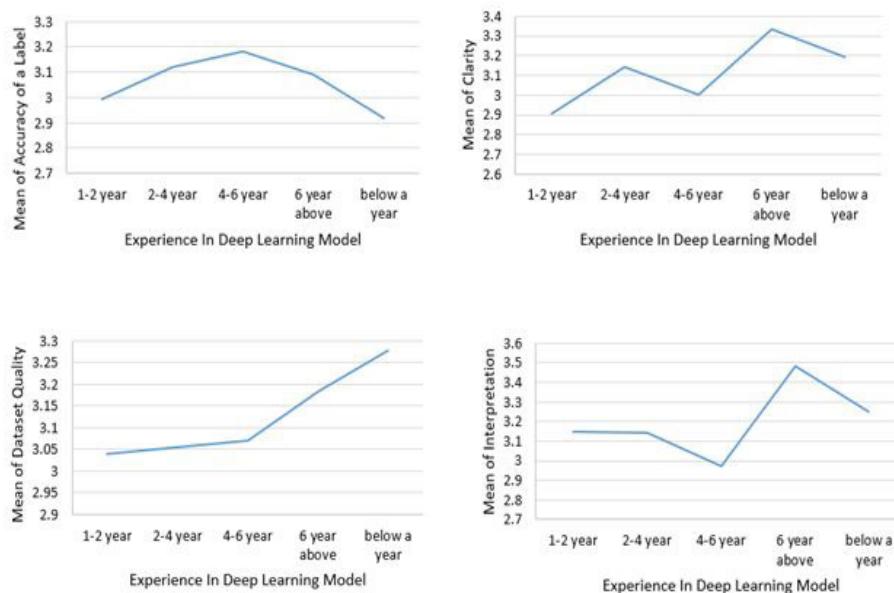


Figure 2: One-Way ANOVA Showing the Effect of Experience in Deep Learning Models on Performance Metrics

Figure 2 illustrates the one-way ANOVA results examining the effect of experience in deep learning models on various performance metrics. The mean trends shown in the graph exhibit only minor variations across the experience levels. This visual representation confirms the statistical findings and shows that there are no significant differences in performance metrics due to the level of experience.

Table 4. One-Way ANOVA Examining Differences in Performance Metrics Across Deep Learning Models

	Sum of Squares	df	Mean Square	F	Sig.
Accuracy of a Label	19.998	15	1.333	1.241	0.237
Contextual Fit	16.648	15	1.110	0.797	0.682
Clarity	24.821	15	1.655	0.987	0.467
Dataset Quality	20.099	15	1.340	0.814	0.663

Citation: Peram, S. R. (2025). Automated Label Detection and Recommendation System Using Deep Convolution Neural Networks Using ANOVA. International Journal of Artificial Intelligence and Machine Learning, 3(4), 1-8. <https://doi.org/10.55124/jaim.v3i4.294>

specificity	32.754	15	2.184	1.163	0.297
Sensitivity	26.835	15	1.789	1.052	0.400
False Positive Rate	28.858	15	1.924	1.318	0.186
Recall	38.454	15	2.564	1.661	0.055
Interpretation	41.419	15	2.761	1.578	0.075

Table 4 presents the one-way ANOVA results comparing the performance metrics across different deep learning models. All evaluated metrics did not show statistically significant differences between the models, as their p-values are greater than 0.05. This indicates that, in this study, the model type did not significantly affect accuracy, contextual relevance, clarity, or other related performance metrics.

Table 5. One-Way ANOVA Examining Differences in Performance Metrics between Model Outputs

	Sum of Squares	df	Mean Square	F	Sig.
Accuracy of a Label	40.717	1	40.717	40.429	0.000
Contextual Fit	1.577	1	1.577	1.139	0.286
Clarity	5.822	1	5.822	3.491	0.062
Dataset Quality	15.749	1	15.749	9.773	0.002
specificity	0.004	1	0.004	0.002	0.965
Sensitivity	0.100	1	0.100	0.059	0.808
False Positive Rate	0.074	1	0.074	0.050	0.822
Recall	1.048	1	1.048	0.666	0.415
Interpretation	1.639	1	1.639	0.921	0.338

Table 5 shows the one-way ANOVA results comparing performance metrics across different model outputs. Statistically significant differences were found in label accuracy and dataset quality ($p < 0.05$), while no significant differences were observed between the model outputs in contextual relevance, clarity, uniqueness, sensitivity, false positive rate, recall, and interpretability.

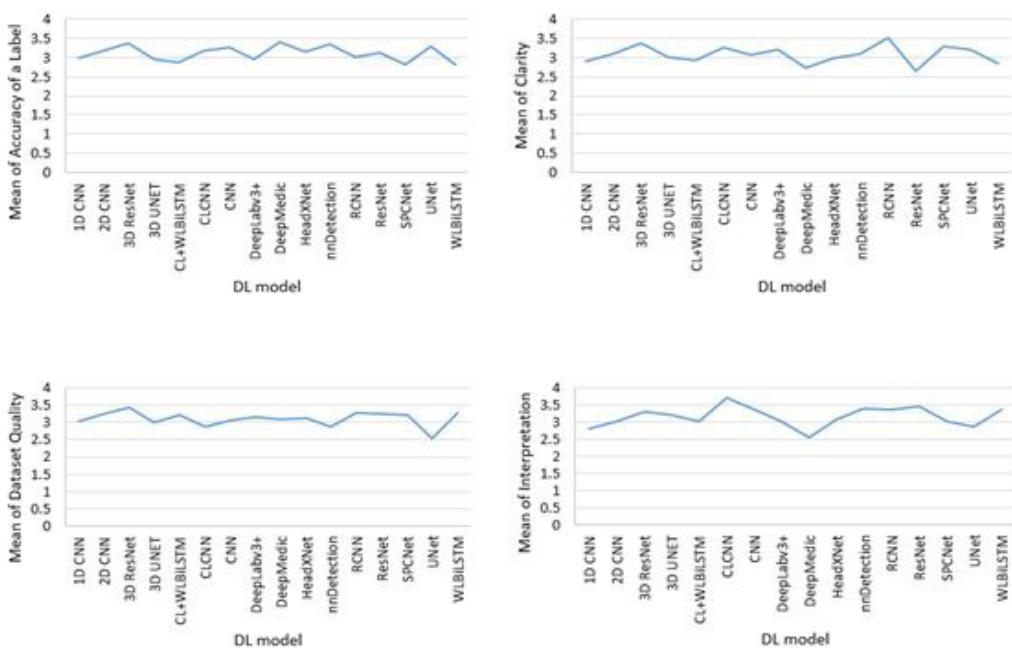


Figure 3: One-Way ANOVA Illustrating Differences in Performance Metrics Across Deep Learning Models

Figure 3 shows one-way ANOVA plots comparing performance metrics across different deep learning models. The mean values across the models are relatively consistent, with no significant deviations. This suggests that the differences between the deep learning architectures did not cause statistically significant variations in the evaluated performance metrics.

Table 6. Levene's Test for Homogeneity of Error Variances for Accuracy of a Label

		Levene Statistic	df1	df2	Sig.
Label Accuracy	Using Mean	1.234	65	455	0.116
	Using Median	0.848	65	455	0.792
	Using Median with adjusted degrees of freedom	0.848	65	338.13	0.788
	Using Trimmed Mean	1.219	65	455	0.129

Evaluates the null hypothesis that error variance for the dependent variable is homogeneous across groups.

a Dependent variable: Label Accuracy

b Design: Intercept + Experience + Deep Learning Model + Experience * Deep Learning Model

Table 6 presents the results of Levene's test, which assesses the homogeneity of variances for label accuracy. Since all significance values exceeded 0.05, it confirms equal variances among the data groups, thus fulfilling the variance assumption required for ANOVA analysis.

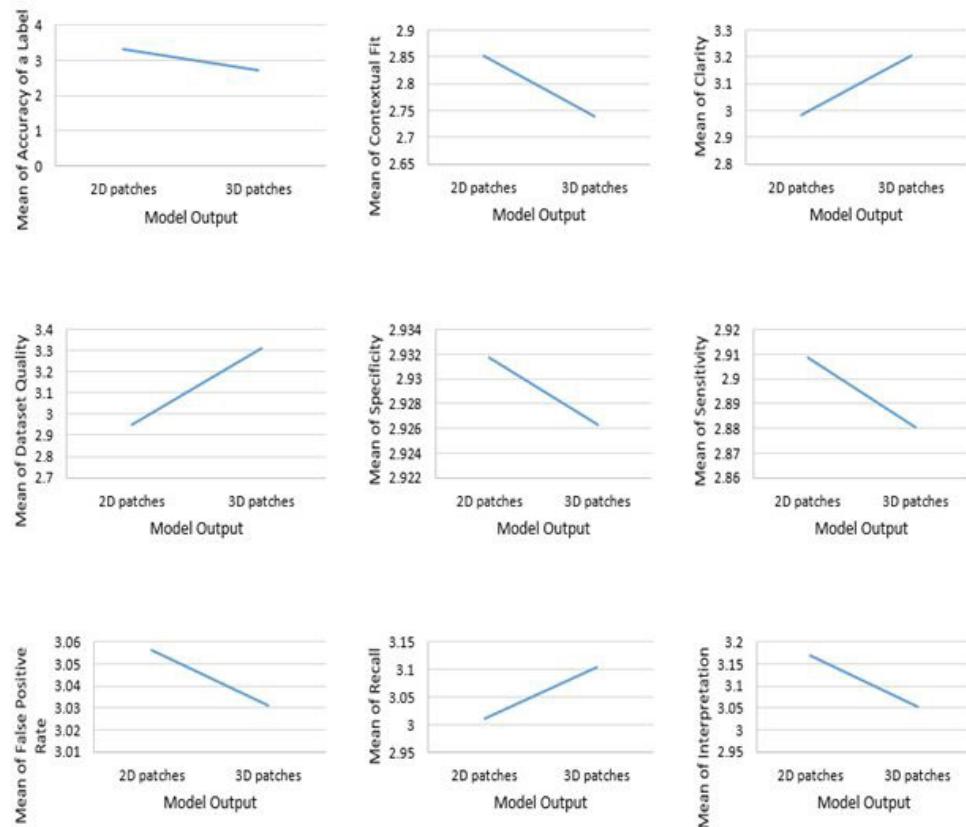


Figure 4: One-Way ANOVA Showing the Effect of Model Output (2D vs 3D Patches) on Performance Measures

Figure 4 shows the one-way ANOVA results comparing the model outputs based on 2D and 3D patches. Significant differences are observed in the accuracy of one label and the quality of the dataset, while other metrics show minimal variation. This indicates that the type of model output significantly affects the chosen performance metrics.

Table 7. Tests of Between-Subjects Effects for Accuracy of a Label

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	98.287a	76	1.293	1.236	0.101	0.171
Intercept	2104.874	1	2104.874	2011.168	0	0.816
Experience	3.808	4	0.952	0.91	0.458	0.008
Deep Learning Model	16.547	15	1.103	1.054	0.398	0.034
Experience * Deep Learning Model	74.751	57	1.311	1.253	0.112	0.136
Error	476.2	455	1.047			
Total	5661	532				
Corrected Total	574.487	531				

a R Squared = .171 (Adjusted R Squared = .033)

Table 7 shows the two-way ANOVA results for label accuracy. It reveals that there are no statistically significant effects from experience, the deep learning model, or their interaction ($p > 0.05$). This model accounts for 17.1% of the total variance, indicating a moderate, but not statistically significant, explanatory power.

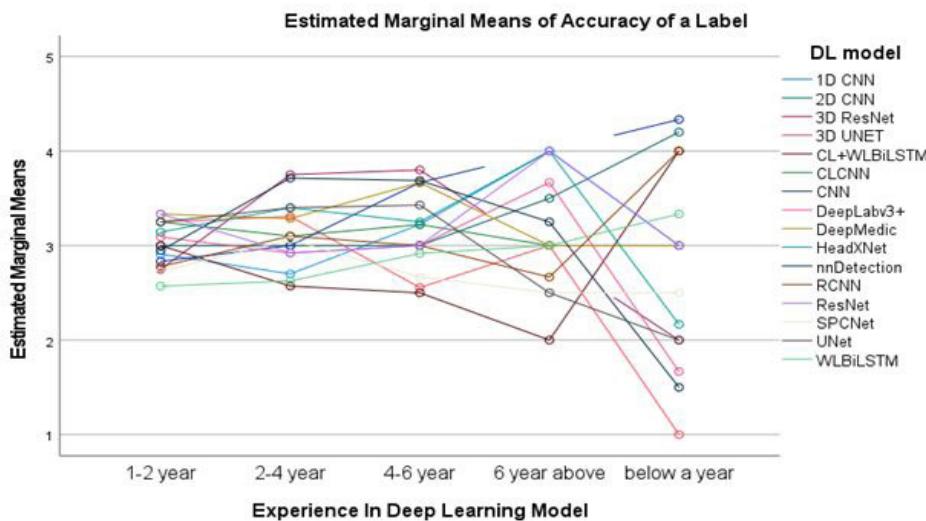


Figure 5: Interaction Effects of Experience in Deep Learning and Deep Learning Model on Accuracy of a Label (Two-Way ANOVA)

Based on the two-way ANOVA results, Figure 4 illustrates how experience in deep learning and the type of deep learning model interact to affect label accuracy. The non-parallel trends indicate variations between the different combinations; however, the absence of strong crossover patterns is consistent with the statistical results showing no significant interaction effect. According to this analysis, the socio-demographic characteristics of the respondents, prior deep learning experience, and model selection did not significantly affect the perceived performance of automated labeling systems. Perceptions of performance were consistent across all groups, while the model output format (2D versus 3D) emerged as a key factor influencing label accuracy and the quality of the dataset.

Conclusion

Automated label detection and recommendation system using deep convolutional neural networks: How deep learning expertise, model selection, and output format affect user evaluations of automated labeling system performance. The results demonstrate that the level of deep learning experience does not meaningfully impact perceptions of accuracy, contextual relevance, clarity, or interpretability, suggesting that contemporary labeling systems provide consistent performance regardless of user expertise. Additionally, differences between deep learning models did not yield statistically significant differences in perceived performance metrics, indicating comparable practical performance across distinct architectures. Model output format emerged as a crucial factor. Significant differences in label accuracy and dataset quality were observed between 2D and 3D patch-based outputs,

highlighting the importance of output representation in shaping system performance. This suggests that output design decisions can significantly impact labeling reliability, regardless of model architecture or user expertise. A two-way ANOVA analysis further confirmed the absence of significant interaction effects between experience and model type on labeling accuracy, reinforcing the robustness of automated labeling systems across diverse user groups. Overall, the study underscores the need to prioritize data representation and output strategies alongside model development. These findings provide crucial guidance for developing efficient and understandable automated coding systems that perform consistently across different user groups.

References

1. Lin, Tian, Meng Wang, Aidi Lin, Xiaoting Mai, Huiyu Liang, Yih-Chung Tham, and Haoyu Chen. "Efficiency and safety of automated label cleaning on multimodal retinal images." *npj Digital Medicine* 8, no. 1 (2025): 10.
2. Kurnaz, Fatih Can, Burak Hocaoglu, Mert Kaan Yilmaz, İdil Sülo, and Sinan Kalkan. "Alet (automated labeling of equipment and tools): A dataset for tool detection and human worker safety detection." In *European Conference on Computer Vision*, pp. 371-386. Cham: Springer International Publishing, 2020.
3. Di Noto, Tommaso, Guillaume Marie, Sébastien Tourbier, Yasser Aleman-Gómez, Oscar Esteban, Guillaume Saliou, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. "Towards automated brain aneurysm detection in TOF-MRA: open data, weak labels, and anatomical knowledge." *Neuroinformatics* 21, no. 1 (2023): 21-34.
4. Hussein, O., Shymaa Mohammed Jameel, J. M. Altmemi, Mohammad A. Abbas, Abbas Uğurenver, Yasir Mahmood Alkubaisi, and Ahmad H. Sabry. "Improving automated labeling with deep learning and signal segmentation for accurate ECG signal analysis." *Service Oriented Computing and Applications* (2024): 1-14.
5. Fang, Weili, Lieyun Ding, Botao Zhong, Peter ED Love, and Hanbin Luo. "Automated detection of workers and heavy equipment on construction sites: A convolutional neural network approach." *Advanced Engineering Informatics* 37 (2018): 139-149.
6. Fanaee-T, Hadi, and Joao Gama. "Event labeling combining ensemble detectors and background knowledge." *Progress in Artificial Intelligence* 2, no. 2 (2014): 113-127.
7. Dakka, M. A., T. V. Nguyen, J. M. M. Hall, S. M. Diakiw, M. VerMilyea, R. Linke, M. Perugini, and D. Perugini. "Automated detection of poor-quality data: case studies in healthcare." *Scientific Reports* 11, no. 1 (2021): 18005.
8. Cui, Yue, Wei Wen, Darren M. Lipnicki, Mirza Faisal Beg, Jesse S. Jin, Suhuai Luo, Wanlin Zhu et al. "Automated detection of amnestic mild cognitive impairment in community-dwelling elderly adults: a combined spatial atrophy and white matter alteration approach." *Neuroimage* 59, no. 2 (2012): 1209-1217.
9. Seetharaman, Arun, Indrani Bhattacharya, Leo C. Chen, Christian A. Kunder, Wei Shao, Simon JC Soerensen, Jeffrey B. Wang et al. "Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging." *Medical Physics* 48, no. 6 (2021): 2960-2972.
10. Leiria, Daniel, Kamilla Heimar Andersen, Simon Pommerenke Melgaard, Hicham Johra, Anna Marszal-Pomianowska, Marco Savino Piscitelli, Alfonso Capozzoli, and Michal Zbigniew Pomianowski. "Towards automated fault detection and diagnosis in district heating customers: generation and analysis of a labeled dataset with ground truth." In *Building Simulation 2023*, vol. 18, pp. 3615-3623. IBPSA, 2023.
11. Cai, Jie, Bin Li, Tao Zhang, Jiale Zhang, and Xiaobing Sun. "Fine-grained smart contract vulnerability detection by heterogeneous code feature learning and automated dataset construction." *Journal of Systems and Software* 209 (2024): 111919.
12. Royston, Sam, Ben Greenberg, Omeed Tavasoli, and Courtenay Cotton. "Anomaly detection and automated labeling for voter registration file changes." *arXiv preprint arXiv:2106.15285* (2021).
13. Brand, Yonatan E., Felix Kluge, Luca Palmerini, Anisoara Paraschiv-Ionescu, Clemens Becker, Andrea Cereatti, Walter Maetzler et al. "Self-supervised learning of wrist-worn daily living accelerometer data improves the automated detection of gait in older adults." *Scientific Reports* 14, no. 1 (2024): 20854.
14. Farrell, Sean, Charlotte Appleton, Peter-John Mäntylä Noble, and Noura Al Moubayed. "PetBERT: automated ICD-11 syndromic disease coding for outbreak detection in first opinion veterinary electronic health records." *Scientific reports* 13, no. 1 (2023): 18015.
15. Lancaster, Jack L., Marty G. Woldorff, Lawrence M. Parsons, Mario Liotti, Catarina S. Freitas, Lacy Rainey, Peter V. Kochunov, Dan Nickerson, Shawn A. Mikiten, and Peter T. Fox. "Automated Talairach atlas labels for functional brain mapping." *Human brain mapping* 10, no. 3 (2000): 120-131.
16. Enkhsaikhan, Majigsuren, Wei Liu, Eun-Jung Holden, and Paul Duuring. "Auto-labelling entities in low-resource text: a geological case study." *Knowledge and Information Systems* 63, no. 3 (2021): 695-715.
17. Rahayu, Nur Indri, M. Muktiarni, and Yusuf Hidayat. "An application of statistical testing: A guide to basic parametric statistics in educational research using SPSS." *ASEAN Journal of Science and Engineering* 4, no. 3 (2024): 569-582.
18. Prasanth, Vidhya. "Developing Business Services Using IBM SPSS Statistics." *REST Journal on Banking, Accounting and Business*.
19. Peram, S. R. "Automated Label Detection and Recommendation System Using Deep Convolution Neural Networks and SPSS-Based Evaluation." *International Journal of Computer Science and Data Engineering* 1, no. 2 (2024): 258.
20. Xiang, Yong, Chenxin Yang, Zhigang Jin, and Wanshu Zhao. "Factors influencing the adoption of generative artificial intelligence into classroom teaching by university teachers: An empirical study using SPSS PROCESS macros." *Plos one* 20, no. 8 (2025): e0324875.