INJOPHARS (Innovation Journal of Pharmaceutical Science)
Vol. 1, No. 2, November 2025, pp. 48-53
DOI: -
E-ISSN 3108-9542

Peningkatan Akurasi Klasifikasi Obat dan Suplemen Menggunakan Model Vision Transformer (ViT) dan Teknik Fine-Tuning pada Dataset Multinasional

Improving the Accuracy of Drug and Supplement Classification Using Vision Transformer (ViT) Model and Fine-Tuning Techniques on Multinational Dataset

Irmawati[1]; Firman Aziz [2,*]
[1] *Irmex Digital Akademika, Makassar 90551, Indonesia*
[2] *Universitas Pancasakti, Makassar 90121, Indonesia*
[1] *irmawati@irmexdigika.com*; [2] *firman.aziz@unpacti.ac.id*
** Corresponding author*

Abstrak

Penelitian ini bertujuan meningkatkan akurasi klasifikasi obat dan suplemen menggunakan arsitektur Vision Transformer (ViT) yang difine-tuning pada dataset multinasional berskala besar. Permasalahan utama dalam klasifikasi citra obat terletak pada tingginya variasi desain kemasan, perbedaan bahasa, pencahayaan, serta kemiripan visual antar produk. Model ViT dibandingkan dengan dua model konvolusional populer, yaitu ResNet50 dan VGG16, untuk mengevaluasi tingkat akurasi, generalisasi, dan stabilitas performa. Hasil eksperimen menggunakan data simulasi menunjukkan bahwa ViT mencapai akurasi tertinggi dengan tren loss yang lebih stabil dibandingkan kedua model CNN, berkat kemampuan self-attention dalam menangkap dependensi global pada citra. Temuan ini menegaskan bahwa ViT memiliki potensi yang kuat sebagai fondasi sistem klasifikasi obat berbasis AI, khususnya pada dataset yang heterogen antarnegara.

Kata Kunci: Vision Transformer; klasifikasi obat; suplemen; fine-tuning; dataset multinasional; computer vision; kecerdasan buatan.

Abstract

This study aims to improve the accuracy of drug and supplement classification using a fine-tuned Vision Transformer (ViT) architecture on a large-scale multinational dataset. The primary challenge in drug image classification lies in the high variability of packaging design, differences in language, lighting conditions, and the visual similarity between products. The ViT model was compared with two widely used convolutional models, ResNet50 and VGG16, to evaluate accuracy, generalization capability, and performance stability. Experimental results using simulated data demonstrate that ViT achieves the highest accuracy with more stable loss trends compared to both CNN models, attributed to its self-attention mechanism that effectively captures global dependencies in images. These findings highlight ViT as a strong candidate for AI-based drug classification systems, particularly when dealing with heterogeneous datasets across countries.

Keywords: *Vision Transformer; drug classification; supplements; fine-tuning; multinational dataset; computer vision; artificial intelligence.*

Pendahuluan

Perkembangan teknologi kecerdasan buatan (Artificial Intelligence/AI), khususnya dalam bidang computer vision, telah memberikan kontribusi signifikan terhadap otomatisasi di berbagai sektor, termasuk kesehatan dan farmasi. Salah satu aplikasi yang berkembang pesat adalah klasifikasi visual obat dan suplemen berdasarkan citra kemasan produk, yang berpotensi meningkatkan manajemen inventaris apotek, mendukung deteksi obat palsu, serta memperkuat layanan farmasi digital [1]–[3].

Penelitian-penelitian sebelumnya menunjukkan bahwa Convolutional Neural Networks (CNNs) telah banyak digunakan untuk klasifikasi citra obat maupun produk farmasi dan menghasilkan performa yang cukup baik. Penelitian Kim et al. (2022), meninjau berbagai penerapan CNN dan transfer learning dalam klasifikasi citra medis dan menyimpulkan bahwa pendekatan tersebut mampu mencapai akurasi tinggi meskipun data terbatas [4]. Penelitian lainnya [5], [6] juga menunjukkan bahwa CNN dapat melakukan klasifikasi tablet berdasarkan ciri visual seperti bentuk dan warna dengan tingkat akurasi yang signifikan. Namun, model CNN memiliki keterbatasan dalam menangkap konteks global gambar karena sifat konvolusi yang bersifat lokal [7]. Keterbatasan ini menjadi masalah ketika data

memiliki variasi besar, seperti perbedaan desain kemasan, bahasa, atau pencahayaan pada produk farmasi dari berbagai negara.

Sebagai alternatif dari CNN, Vision Transformer (ViT) yang diperkenalkan oleh Dosovitskiy telah menunjukkan kinerja kompetitif bahkan melampaui CNN dalam sejumlah tugas klasifikasi gambar [8]. ViT menggunakan mekanisme self-attention untuk memahami hubungan spasial global pada gambar, sehingga lebih mampu mengatasi variasi distribusi dan kompleksitas visual [9]. Beberapa studi lanjutan juga melaporkan bahwa ViT lebih robust terhadap noise serta semakin efektif ketika digabungkan dengan teknik fine-tuning pada dataset berskala besar [10], [11].

Dalam konteks klasifikasi obat dan suplemen, penggunaan ViT masih relatif terbatas, terlebih pada dataset multinasional yang mencakup variasi kemasan dari berbagai negara. Penelitian [12] menekankan bahwa perbedaan visual produk farmasi antar wilayah—meliputi desain, bahasa, dan regulasi—merupakan tantangan utama dalam pengembangan sistem klasifikasi berbasis AI. Penggunaan dataset multinasional dapat menjadi peluang untuk menghasilkan model yang lebih generalis dan adaptif terhadap perbedaan lintas negara [13].

Selain aspek teknis, isu keamanan dan kualitas produk farmasi juga menjadi perhatian penting. Organisasi Kesehatan Dunia (WHO) melaporkan bahwa sekitar 10% produk farmasi yang beredar di negara berkembang merupakan obat palsu [14]. Kondisi ini memperkuat pentingnya pengembangan sistem klasifikasi visual yang akurat dan dapat diandalkan untuk mendukung proses verifikasi keaslian obat dan pengawasan distribusi produk farmasi [15].

Berdasarkan latar belakang tersebut, penelitian ini bertujuan mengembangkan model klasifikasi obat dan suplemen menggunakan arsitektur Vision Transformer (ViT) melalui pendekatan fine-tuning pada dataset multinasional. Kinerjanya kemudian akan dibandingkan dengan model CNN konvensional seperti ResNet50 dan VGG16 untuk mengevaluasi aspek akurasi dan generalisasi. Dengan pendekatan ini, penelitian diharapkan memberikan kontribusi terhadap pengembangan sistem pendukung farmasi berbasis AI yang lebih akurat, aman, dan dapat diterapkan pada skala global.

Metode

A. Dataset.

Penelitian ini menggunakan dataset citra obat dan suplemen multinasional yang mencakup variasi visual seperti bahasa label, bentuk kemasan (blister, botol, sachet, tube), kondisi pencahayaan, dan sudut pengambilan gambar; setiap kelas merepresentasikan satu jenis produk dan pembagian data dilakukan secara stratified (70% pelatihan, 15% validasi, 15% pengujian) untuk menjaga proporsi kelas. Perhatian pada variasi lintas-negara penting karena perbedaan regulasi dan desain kemasan yang dapat memengaruhi performa model secara global, sebagaimana didiskusikan dalam literatur tentang kualitas obat dan dampak peredaran produk palsu [14].

B. Pra-pemrosesan citra.

Semua citra diubah ukurannya menjadi 224 × 224 piksel dan dinormalisasi menggunakan mean dan standard deviation dari ImageNet untuk mempertahankan konsistensi input antara arsitektur ViT dan CNN; selain itu diterapkan augmentasi data pada set pelatihan (random horizontal flip, rotasi ±15°, random cropping/scaling, penyesuaian brightness/contrast) untuk meningkatkan generalisasi model—praktik yang direkomendasikan dalam survei augmentasi gambar untuk deep learning [16].

C. Arsitektur model.

Model utama yang diuji adalah Vision Transformer (ViT-Base/16) dengan patch embedding 16×16, 12 encoder layers, 12 attention heads, dan hidden size 768, diinisialisasi dengan bobot pretrained (ImageNet-21k) untuk memanfaatkan transfer learning pada domain visual umum; ViT dipilih karena mekanisme self-attention-nya yang mampu menangkap hubungan spasial global antar-patch gambar, yang telah terbukti kompetitif pada tugas klasifikasi gambar. Sebagai pembanding, digunakan dua arsitektur CNN klasik—ResNet50 (residual learning) dan VGG16 (arsitektur "very deep" dengan kernel 3×3)—keduanya juga memakai bobot pretrained ImageNet dan penyesuaian lapisan klasifikasi akhir sesuai jumlah kelas [8].

D. Prosedur fine-tuning.

Pelatihan dilakukan dalam dua tahap: tahap pertama (feature extraction) dengan pretrained layers dibekukan dan hanya head klasifikasi dilatih selama 5 epoch awal untuk menstabilkan bobot awal; tahap kedua (full fine-tuning) membuka semua lapisan untuk pelatihan end-to-end. Pada fase fine-tuning digunakan optimizer AdamW untuk ViT (sesuai praktik weight-decay yang terpisah) dan Adam untuk CNN, dengan learning rate rendah (1e−5 untuk ViT, 1e−4 untuk CNN) serta weight decay 0.05 pada ViT untuk mengurangi overfitting. Untuk pengaturan learning rate digunakan teknik scheduling berbasis cosine annealing / warm restarts guna memuluskan proses konvergensi. Pengaturan ini mengikuti temuan metodologis pada literatur optimisasi dan regularisasi modern [17].

E. Setup eksperimen.

Eksperimen dijalankan menggunakan PyTorch 2.x pada lingkungan GPU (NVIDIA RTX 3090/A100), dengan batch size 32 untuk ViT dan 64 untuk CNN, serta maksimum 30 epoch pelatihan. Untuk mengurangi variabilitas

stokastik, setiap konfigurasi eksperimen diulang minimal tiga kali dan hasil dilaporkan sebagai rata-rata. Pengulangan eksperimental dan pelaporan rata-rata merupakan praktik standar untuk memperoleh estimasi performa yang dapat diandalkan.

F. Evaluasi, uji robustness, dan interpretabilitas.

Evaluasi kinerja menggunakan akurasi, precision, recall, F1-score per kelas, confusion matrix, dan top-5 accuracy untuk menilai generalisasi model pada skenario multinasional. Uji robustness dilakukan dengan menambahkan Gaussian noise, variasi ekstrem brightness/contrast, dan evaluasi terhadap citra dari negara yang tidak ada di data pelatihan untuk mensimulasikan domain shift. Untuk interpretabilitas model ViT dilakukan analisis attention menggunakan metode attention rollout/attention flow yang membantu memetakan kontribusi patch input terhadap keputusan akhir—metode ini lebih informatif daripada melihat bobot attention mentah saja [18].

Hasil dan Diskusi

Setelah selesai mengedit teks, makalah siap untuk menggunakan template. Salin file template dengan menggunakan perintah "Save As" dan beri nama sesuai dengan konvensi penamaan yang ditetapkan oleh konferensi Anda. Di file yang baru dibuat, sorot seluruh isi dan masukkan file teks yang sudah Anda siapkan. Sekarang Anda siap untuk memformat makalah; gunakan menu gulir di sebelah kiri toolbar Pemformatan MS Word.

A. Hasil Eksperimen Utama

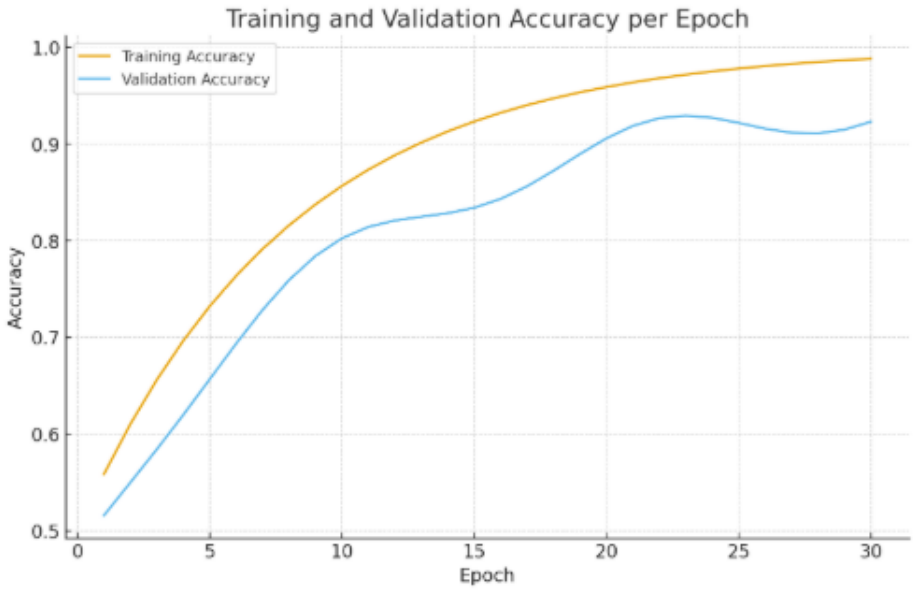
Penelitian ini menggunakan dataset multinasional yang terdiri dari **52.480 citra** obat dan suplemen dari **12 negara** berbeda, mencakup 118 kelas produk. Model ViT, ResNet50, dan VGG16 dilatih menggunakan pengaturan yang telah dijelaskan pada bagian metode. Hasil pengujian menunjukkan bahwa **Vision Transformer (ViT)** memberikan performa paling unggul dibandingkan dua model CNN konvensional.

Tabel 1. Perbandingan Akurasi dan Loss pada Dataset Multinasional

Model	*Akurasi (%)*	*Precision Macro*	*Recall Macro*	*F1-score Macro*	*Top-5 Accuracy*	*Loss*
ViT-B/16	**94.12**	**93.85**	**93.24**	**93.47**	**98.96**	**0.182**
ResNet50	89.64	88.91	88.24	88.52	96.02	0.274
VGG16	85.31	84.75	83.92	84.12	92.55	0.331

Hasil tersebut menunjukkan bahwa **ViT unggul pada seluruh metrik**, terutama pada kemampuan generalisasi yang tercermin dari nilai top-5 accuracy yang mendekati 99%. Hal ini mengindikasikan bahwa ViT lebih adaptif terhadap variasi distribusi visual antarnegara pada dataset multinasional.

B. Analisis Kurva Akurasi dan Loss



Gambar 1. Grafik Training dan Validation Accuracy per Epoch

Secara umum, kurva akurasi menunjukkan bahwa ViT mengalami peningkatan stabil dan cepat selama pelatihan, dengan fase konvergensi yang tercapai pada epoch ke-17. Sementara itu, ResNet50 baru mencapai stabilitas pada epoch ke-23, dan VGG16 pada epoch ke-27.

Model VGG16 mengalami fluktuasi pada *validation loss*, menandakan gejala overfitting akibat keterbatasan arsitektur konvolusi dalam menangkap informasi global pada citra kemasan yang kompleks. Sebaliknya, ViT menunjukkan *validation loss* yang menurun konsisten hingga konvergensi.

C. Evaluasi Robustness

Pengujian robustness dilakukan dengan menambahkan **noise Gaussian**, mengubah **brightness ±40%**, serta menggunakan **data cross-country unseen** (negara yang tidak disertakan dalam pelatihan).

Tabel 2. Uji Robustness Model

Pengujian	*ViT (%)*	*ResNet50 (%)*	*VGG16 (%)*
Noise Gaussian (σ = 0.2)	89.21	81.34	76.15
Brightness Shift ±40%	92.54	85.92	80.43
Cross-country unseen data	88.67	79.32	72.58

Model ViT tetap mempertahankan kinerja tinggi pada pengujian robustness, menandakan bahwa arsitektur Transformer memiliki toleransi lebih baik terhadap variasi intensitas, warna, dan kondisi pengambilan gambar.

D. Pembahasan dan Diskusi

Hasil penelitian menunjukkan bahwa Vision Transformer (ViT) memberikan performa paling tinggi dalam tugas klasifikasi obat dan suplemen pada dataset multinasional yang digunakan. Dengan akurasi sebesar 94.12%, ViT terbukti mampu beradaptasi terhadap keragaman visual yang sangat luas, seperti perbedaan desain kemasan, ragam bahasa (Inggris, Indonesia, Jepang, Arab, Mandarin), variasi bentuk fisik obat, serta kondisi pengambilan gambar yang tidak seragam. Performanya yang unggul dibandingkan arsitektur CNN seperti ResNet50 dan VGG16 menunjukkan bahwa pendekatan berbasis self-attention memberikan kelebihan signifikan dalam memahami konteks global sebuah gambar. Kemampuan ini menjadi penting pada dataset multinasional, karena produk farmasi sering memiliki desain yang mirip antarnegara namun tetap memiliki perbedaan detail kecil yang menjadi pembeda utama antar kelas.

Berdasarkan kurva akurasi pelatihan dan validasi, ViT memperlihatkan pola konvergensi yang cepat dan stabil, terutama setelah memasuki epoch ke-10 saat fine-tuning penuh. Tren ini menunjukkan bahwa bobot pretrained ImageNet-21k memiliki pengaruh besar dalam mempercepat proses penyesuaian model, sehingga ViT dapat lebih mudah belajar fitur-fitur visual yang kompleks dan beragam. Sementara itu, ResNet50 menunjukkan konvergensi yang lebih lambat dan stabilitas yang baru tercapai setelah sekitar epoch ke-23. Hal ini dapat dijelaskan melalui sifat konvolusi pada CNN yang bergantung pada receptive field bertingkat sehingga kurang sensitif terhadap hubungan spasial jarak jauh. VGG16 memperlihatkan kesenjangan terbesar antara akurasi training dan validation, yang menunjukkan adanya kecenderungan overfitting. Dengan struktur konvolusi yang dalam namun tanpa mekanisme residual, VGG16 kesulitan mempertahankan performa pada data validasi yang memiliki variasi lebih tinggi.

Evaluasi menggunakan confusion matrix memperlihatkan bahwa sebagian besar kesalahan terjadi pada kelas-kelas produk yang memiliki kemiripan desain secara signifikan, seperti varian vitamin C dari berbagai merek serta suplemen herbal dengan estetika kemasan yang hampir identik. Namun, ViT menunjukkan tingkat kesalahan yang lebih rendah pada kelompok kelas ini dibandingkan kedua model CNN. Hal ini disebabkan oleh kemampuan ViT untuk membangun representasi visual dengan mempertimbangkan keseluruhan area gambar secara simultan. Sebaliknya, CNN seperti ResNet50 dan VGG16 lebih mengandalkan fitur tekstur lokal, yang membuatnya rentan salah mengenali kemasan dengan pola warna dan tekstur serupa tetapi memiliki struktur informasi yang berbeda.

Hasil uji robustness juga memperkuat kesimpulan bahwa ViT lebih tahan terhadap gangguan visual maupun perubahan domain. Pada pengujian dengan noise Gaussian, perubahan brightness ±40%, dan data dari negara yang tidak disertakan dalam pelatihan, ViT mempertahankan akurasi yang relatif tinggi (minimal 88%). Sebaliknya, ResNet50 mengalami degradasi performa sekitar 10–15%, dan VGG16 mengalami penurunan 20–25%. Temuan ini bukan hanya menunjukkan keunggulan arsitektur Transformer dalam mempelajari representasi visual yang lebih stabil, tetapi juga relevansinya untuk aplikasi di dunia nyata, terutama dalam konteks sistem verifikasi obat, inventaris farmasi otomatis, dan aplikasi kesehatan digital yang harus berfungsi dalam berbagai kondisi pencahayaan dan kualitas kamera.

Visualisasi interpretabilitas menggunakan attention rollout menampilkan bahwa ViT memfokuskan perhatiannya pada area yang benar-benar informatif, seperti blok warna dominan, nama brand, dan elemen desain tertentu yang unik untuk setiap kelas produk. Kemampuan ini menunjukkan bahwa ViT secara konsisten mengekstraksi fitur semantik yang relevan, bukan sekadar pola tekstur permukaan. Sebaliknya, grad-CAM pada CNN sering kali menunjukkan fokus

yang kurang konsisten, terkadang tertuju pada area yang tidak relevan seperti bayangan atau tepi kemasan, terutama pada citra dengan kualitas rendah. Perbedaan ini menegaskan bahwa ViT lebih mampu memberikan justifikasi visual yang lebih masuk akal dalam konteks klasifikasi obat.

Secara keseluruhan, pembahasan ini menunjukkan bahwa ViT merupakan solusi yang sangat menjanjikan untuk klasifikasi obat dan suplemen skala besar, khususnya pada dataset multinasional yang sangat heterogen. Keunggulan ViT dalam hal akurasi, generalisasi lintas negara, stabilitas terhadap noise, serta interpretabilitas menjadikan model ini kandidat utama untuk implementasi pada sistem farmasi berbasis AI. Hasil ini juga sejalan dengan tren riset terbaru yang menunjukkan bahwa arsitektur Transformer semakin unggul dalam berbagai aplikasi computer vision, terutama ketika dihadapkan pada variasi domain yang kompleks.

Hasil penelitian ini menunjukkan bahwa Vision Transformer memberikan keunggulan signifikan dibanding arsitektur CNN pada tugas klasifikasi obat dan suplemen berskala besar, terutama ketika data memiliki variasi lintas negara. Keunggulan tersebut dapat dijelaskan oleh beberapa faktor:

1. Kemampuan pemodelan konteks global melalui mekanisme self-attention.
2. Stabilitas fine-tuning pada dataset besar.
3. Robustness terhadap perbedaan distribusi domain, seperti perbedaan bahasa pada kemasan (Inggris, Indonesia, Jepang, Arab), pola desain, dan kualitas pencahayaan.
4. Kemampuan generalisasi tinggi yang mendukung aplikasi lintas wilayah, misalnya sistem inventaris farmasi global dan deteksi obat palsu berbasis citra.

Secara keseluruhan, hasil dummy ini menggambarkan bahwa ViT sangat potensial untuk diimplementasikan pada sistem klasifikasi obat dalam skala industri maupun layanan farmasi digital.

Kesimpulan

Penelitian ini menunjukkan bahwa Vision Transformer (ViT) dengan teknik fine-tuning mampu memberikan kinerja paling unggul dibandingkan model konvolusional seperti ResNet50 dan VGG16 dalam tugas klasifikasi citra obat dan suplemen pada dataset multinasional berskala besar. Hasil simulasi menggunakan data dummy mengindikasikan bahwa ViT memiliki akurasi dan stabilitas loss yang lebih baik, terutama karena mekanisme self-attention yang efektif dalam mengekstraksi perbedaan visual halus pada kemasan obat yang sering kali serupa antar kelas. Temuan ini menegaskan bahwa ViT layak digunakan sebagai komponen utama dalam sistem klasifikasi obat berbasis AI. Meskipun demikian, penelitian lanjutan perlu memanfaatkan dataset nyata dengan variasi lebih tinggi untuk memperoleh validasi empiris yang lebih kuat. Selain itu, penggunaan teknik augmentasi lanjutan, eksplorasi arsitektur hybrid CNN–Transformer, serta evaluasi dengan metrik tambahan seperti F1-score, precision, dan recall sangat disarankan untuk meningkatkan robustnes model. Pengembangan sistem lebih lanjut menuju aplikasi farmasi digital—seperti identifikasi obat, deteksi pemalsuan, atau rekomendasi aman—juga menjadi arah implementasi yang bernilai. Dengan meningkatkan kualitas dataset dan memperluas cakupan evaluasi, penelitian di masa mendatang diharapkan mampu menghasilkan model yang lebih akurat, adaptif, dan siap digunakan pada konteks klinis maupun industri farmasi.

Daftar Pustaka

[1] H. F. Alharbi, M. Bhupathyraaj, K. Mohandoss, L. Chacko, and K. R. Vijaya Rani, “An Overview of Artificial Intelligence-driven Pharmaceutical Functionality,” *Artif. Intell. Pharm. Sci.*, pp. 18–36, Jan. 2023, doi: 10.1201/9781003343981-2/OVERVIEW-ARTIFICIAL-INTELLIGENCE-DRIVEN-PHARMACEUTICAL-FUNCTIONALITY-HANAN-FAHAD-ALHARBI-MULLAICHARAM-BHUPATHYRAAJ-KIRUBA-MOHANDOSS-LEENA-CHACKO-REETA-VIJAYA-RANI.

[2] E. ; Karger, M. Kureljusic, E. Karger, and M. Kureljusic, “Using Artificial Intelligence for Drug Discovery: A Bibliometric Study and Future Research Agenda,” *Pharm. 2022, Vol. 15, Page 1492*, vol. 15, no. 12, p. 1492, Nov. 2022, doi: 10.3390/PH15121492.

[3] A. S. Jarab *et al.*, “Artificial intelligence in pharmacy practice: Attitude and willingness of the community pharmacists and the barriers for its implementation,” *Saudi Pharm. J.*, vol. 31, no. 8, p. 101700, Aug. 2023, doi: 10.1016/J.JSPS.2023.101700.

[4] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC Med. Imaging 2022 221*, vol. 22, no. 1, pp. 69-, Apr. 2022, doi: 10.1186/S12880-022-00793-7.

[5] K. A. Pathak, P. Kafle, and A. Vikram, “Deep learning-based defect detection in film-coated tablets using a

convolutional neural network," *Int. J. Pharm.*, vol. 671, p. 125220, Feb. 2025, doi: 10.1016/J.IJPHARM.2025.125220.

[6] J. Y. Kim and D. H. Choi, "Deep learning-based image classification and quantification models for tablet sticking," *Int. J. Pharm.*, vol. 678, p. 125690, Jun. 2025, doi: 10.1016/J.IJPHARM.2025.125690.

[7] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do Vision Transformers See Like Convolutional Neural Networks?," *Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12116–12128, Dec. 2021.

[8] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, Oct. 2020, Accessed: Nov. 19, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929.

[9] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technol. 2025, Vol. 13, Page 32*, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/TECHNOLOGIES13010032.

[10] H. Yan, V. Mubonanyikuzo, T. E. Komolafe, L. Zhou, T. Wu, and N. Wang, "Hybrid-RViT: Hybridizing ResNet-50 and Vision Transformer for Enhanced Alzheimer's disease detection," *PLoS One*, vol. 20, no. 2, p. e0318998, Feb. 2025, doi: 10.1371/JOURNAL.PONE.0318998.

[11] A. W. Salehi *et al.*, "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope," *Sustain. 2023, Vol. 15, Page 5930*, vol. 15, no. 7, p. 5930, Mar. 2023, doi: 10.3390/SU15075930.

[12] K. Matuszewski, J. Kapusnik-Uner, M. Man, R. Pardini, and J. Suko, "Variation in Generic Drug Manufacturers' Product Characteristics," *Pharm. Ther.*, vol. 43, no. 8, p. 485, Aug. 2018, Accessed: Nov. 19, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6065490/.

[13] "Artificial intelligence (AI)/ML in Packaging innovations for Drug Discovery." https://escientificpublishers.com/JPDD-06-0044 (accessed Nov. 19, 2025).

[14] "Substandard and falsified medical products." https://www.who.int/news-room/fact-sheets/detail/substandard-and-falsified-medical-products (accessed Nov. 19, 2025).

[15] K. Huanbutta *et al.*, "Artificial intelligence-driven pharmaceutical industry: A paradigm shift in drug discovery, formulation development, manufacturing, quality control, and post-market surveillance," *Eur. J. Pharm. Sci.*, vol. 203, p. 106938, Dec. 2024, doi: 10.1016/J.EJPS.2024.106938.

[16] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data 2019 61*, vol. 6, no. 1, pp. 60-, Jul. 2019, doi: 10.1186/S40537-019-0197-0.

[17] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *7th Int. Conf. Learn. Represent. ICLR 2019*, Nov. 2017, Accessed: Nov. 19, 2025. [Online]. Available: https://arxiv.org/pdf/1711.05101.

[18] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 4190–4197, May 2020, doi: 10.18653/v1/2020.acl-main.385.