Nice — I dug into the recent and influential literature and compiled a compact research brief you can use as the backbone of a literature review. Below you’ll find: (A) the seminal/high-impact works, (B) common datasets and benchmarks, (C) dominant model families & techniques used, (D) typical evaluation/validation practices and pitfalls, and (E) short suggestions for where to focus next. I included citations to the most important sources so you can open them directly.

**Quick findings (executive)**

* Deep CNNs trained with large, clinically-annotated image collections were shown to reach dermatologist-level performance (Esteva et al., 2017). ([Nature](https://www.nature.com/articles/nature21056?utm_source=chatgpt.com))
* Public benchmark datasets (ISIC series, HAM10000) and the ISIC Grand Challenge have driven most progress and standardised tasks (segmentation, attribute detection, classification). ([challenge.isic-archive.com](https://challenge.isic-archive.com/?utm_source=chatgpt.com))
* Modern SOTA uses transfer learning on strong backbones (EfficientNet, ResNet variants, ConvNeXt), multimodal inputs, careful augmentation, class-imbalance handling, and increasingly Vision Transformers / hybrid architectures. ([Nature](https://www.nature.com/articles/s41598-024-82241-w?utm_source=chatgpt.com))
* For lesion localization/segmentation, U-Net and variants (attention U-Net, dual-stage pipelines) remain common and effective. ([ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1361841523001238?utm_source=chatgpt.com))
* Reviews and systematic analyses point out frequent issues: dataset bias, limited external validation, label noise, and lack of clinical metadata in training — all important when aiming for deployment. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC8391467/?utm_source=chatgpt.com))

**A. Seminal / high-impact papers & results**

* **Esteva et al., Nature 2017 — “Dermatologist-level classification…”**  
  Trained a CNN on ~129k images across >2000 disease labels; performance comparable to dermatologists for certain binary tasks. This paper popularised pretraining + fine-tuning with large clinical datasets. ([Nature](https://www.nature.com/articles/nature21056?utm_source=chatgpt.com))
* **ISIC Grand Challenge (annual)** — community benchmark that provides tasks (segmentation, classification) and clear leaderboards. Winner approaches have evolved from Ensembles of CNNs to more modern hybrid models and multimodal systems. ([challenge.isic-archive.com](https://challenge.isic-archive.com/?utm_source=chatgpt.com))
* **HAM10000 dataset paper (Tschandl et al., 2018)** — key publicly available dermatoscopic dataset (≈10k images) used widely for multiclass classification research. ([Nature](https://www.nature.com/articles/sdata2018161?utm_source=chatgpt.com))
* **Systematic reviews (2021–2023)** — surveys summarizing architectures, common preprocessing, and pitfalls (dataset shift, imbalance, explainability needs). Useful for understanding methodological trends and risks. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC8391467/?utm_source=chatgpt.com))

**B. Common datasets / benchmarks**

* **ISIC archive & ISIC Challenges (2016–present)** — tasks: segmentation, attribute detection, classification; datasets combine dermoscopy and clinical images and are the de-facto benchmark. ([challenge.isic-archive.com](https://challenge.isic-archive.com/?utm_source=chatgpt.com))
* **HAM10000 (≈10,015 dermatoscopic images)** — multi-source, pathology-verified subset, widely used for multiclass classification (7 classes). ([Nature](https://www.nature.com/articles/sdata2018161?utm_source=chatgpt.com))
* **ISIC 2018 / 2019 collections (and Kaggle aggregations)** — used widely in recent papers & experiments. ([Kaggle](https://www.kaggle.com/datasets/andrewmvd/isic-2019?utm_source=chatgpt.com))

**C. Typical model families & techniques**

* **CNN backbones / transfer learning** — ResNet, Inception, DenseNet, EfficientNet pretrained on ImageNet, then fine-tuned on dermatology data (standard pipeline since Esteva et al.). Ensembles of fine-tuned models are common to raise robustness. ([Nature](https://www.nature.com/articles/nature21056?utm_source=chatgpt.com))
* **Lightweight/mobile models** — MobileNetV2 and lightweight EfficientNet variants are used when targeting smartphone deployment (Esteva paper highlighted mobile potential). ([Computer Science](https://cs.stanford.edu/people/esteva/nature/?utm_source=chatgpt.com))
* **Segmentation models** — U-Net and its extensions (attention U-Net, cascaded/dual-stage) for lesion masks before classification or for aiding explainability. ([ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1361841523001238?utm_source=chatgpt.com))
* **Hybrid / Transformer models** — Vision Transformers (ViT) or hybrid Conv+Self-Attention (ConvNeXt + separable attention, ConvNeXtV2 hybrids) are emerging in 2023–2025 literature with promising gains. ([Nature](https://www.nature.com/articles/s41598-025-89230-7?utm_source=chatgpt.com))
* **Multimodal approaches** — combining clinical metadata (age, sex, lesion location) or multiple image modalities often improves performance vs. image-only models. ([MedRxiv](https://www.medrxiv.org/content/10.1101/2024.05.30.24308213v1?utm_source=chatgpt.com" \o "Accurate Skin Lesion Classification Using Multimodal ...))

**D. Typical training & evaluation practices (and issues)**

* **Data augmentations** (rotation, color jitter, dermoscopy-specific augmentation), class re-weighting, focal loss, oversampling of minority classes are standard to combat class imbalance (many lesion datasets are heavily skewed). ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9497837/?utm_source=chatgpt.com))
* **Cross-validation and external test sets** — best practice is to evaluate on external datasets (different clinic/population) to measure generalization; many papers fall short here. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9841357/?utm_source=chatgpt.com))
* **Explainability** — Grad-CAM and related saliency maps are commonly applied to show where models focus; required for clinical trust but not a full substitute for clinical validation. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC8391467/?utm_source=chatgpt.com))
* **Pitfalls:** label noise, selection bias, overfitting to dataset artifacts (e.g., rulers, skin markers), and lack of prospective clinical testing. Systematic reviews repeatedly call for stronger external validation. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC9841357/?utm_source=chatgpt.com))

**E. Where to focus for your project (practical suggestions)**

1. **Start with established benchmarks**: experiment on HAM10000 + ISIC subset so your results are comparable to the literature. ([Nature](https://www.nature.com/articles/sdata2018161?utm_source=chatgpt.com))
2. **Use transfer learning + strong backbone**: EfficientNet or ConvNeXt for a good accuracy/compute tradeoff; consider ensembled fine-tuned checkpoints for best validation scores. ([Nature](https://www.nature.com/articles/s41598-024-82241-w?utm_source=chatgpt.com))
3. **Add segmentation as support**: train a U-Net lesion segmenter and use its mask as an extra channel or to crop input—this often improves classification. ([ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1361841523001238?utm_source=chatgpt.com))
4. **Address imbalance & label quality**: use class weighting, focal loss or oversampling; if possible, re-verify labels or use pathology-confirmed subsets for evaluation. ([Nature](https://www.nature.com/articles/sdata2018161?utm_source=chatgpt.com))
5. **Explainability & external validation**: include Grad-CAM visualizations and test on at least one external dataset (different source) before making clinical claims. ([PMC](https://pmc.ncbi.nlm.nih.gov/articles/PMC8391467/?utm_source=chatgpt.com))

**Ready-to-use next steps I can do for you (pick any)**

* Produce a **detailed table** (CSV/Excel) of ~12–20 specific papers: title, year, model type, dataset(s) used, key techniques, performance metrics, DOI/URL.
* Crawl and summarise **ISIC Challenge winning methods** for the years you care about (e.g., 2018–2024).
* Build a **reproducible baseline plan**: code + training recipe (backbone, augmentations, losses, hyperparams) you can run on HAM10000.
* Create a short **literature review paragraph** you can paste into a paper/thesis with citations.

Which of those would you like me to do now? If you want the detailed table, I’ll compile the key papers and include links and short annotations.

How we stand out our model :

Great question — that’s exactly what a strong research project should ask. Most existing skin disease detection models already achieve **high accuracy** on benchmark datasets, so to make yours *better*, you should go beyond just accuracy and tackle the weaknesses of prior work.

Here’s a strategy for improvement — both **technical** and **research-novelty-wise**:

**🔑 1. Use Better Data Strategy**

Most papers suffer from **dataset bias** (HAM10000 and ISIC are very popular but have limited diversity).

* **External validation** – test your model on multiple datasets (e.g., HAM10000 + ISIC + Derm7pt). A model that generalizes across datasets is much more valuable.
* **Data augmentation with medical priors** – not just flips/rotations. Use stain normalization, synthetic hair removal, and realistic GAN-based lesion augmentation (e.g., StyleGAN or diffusion models).
* **Metadata inclusion** – use patient age, sex, and lesion location if available. Studies show a boost in performance when combining image + metadata inputs.

**🧠 2. Build a Stronger Architecture**

Most prior work fine-tunes **off-the-shelf CNNs**. You can:

* **Hybrid models** – combine **CNN for local texture + Vision Transformer for global context** (ConvNeXt + ViT).
* **Multi-scale learning** – lesions have fine details; add a multi-scale feature extractor (e.g., feature pyramid network).
* **Attention mechanisms** – channel/spatial attention (CBAM, SE blocks) help focus on lesion regions and reduce background noise.
* **Ensemble learning** – combine predictions of several backbones (e.g., EfficientNet + DenseNet) for more robust predictions.

**🎯 3. Tackle Class Imbalance & Noise**

Most datasets are **imbalanced** (melanoma is rare):

* Use **focal loss** or **class-balanced loss** to prioritize rare diseases.
* **Self-supervised pretraining** – contrastive learning (SimCLR, MoCo) can make feature representations stronger even with limited labels.
* **Label noise handling** – use semi-supervised learning (Noisy Student, pseudo-labeling) to clean noisy labels.

**🩺 4. Improve Explainability & Trust**

A model that’s just accurate is not enough for clinical adoption.

* Use **Grad-CAM++, LIME, SHAP** to highlight lesion areas and give interpretable outputs.
* Pair classification with **lesion segmentation masks** to show *where* the model is focusing.
* Build a **confidence scoring system** (calibrated probabilities) to tell doctors when the model is uncertain.

**🌍 5. Focus on Deployment & Real-World Utility**

Many research papers stop at benchmark results. You can go further:

* Build a **lightweight version** (quantized, pruned) for real-time inference on smartphones.
* Add a **human-in-the-loop** workflow — allow dermatologist feedback to improve future predictions.
* Test in **uncontrolled settings** (different lighting, smartphone photos), which most research models fail on.

**🧪 6. Explore Novel Research Directions**

* **Multimodal fusion** – combine dermoscopic + clinical (macro) images if available.
* **Few-shot learning** – handle very rare diseases with metric learning (prototypical networks).
* **Uncertainty-aware models** – Bayesian deep learning or Monte Carlo dropout for risk estimation.

**🔥 Bottom line: How to Beat Prior Models**

**Better Generalization + Trust + Real-World Usability = Stronger Research Contribution**

Most previous models just optimize accuracy on one dataset.  
If you make a model that:

* Generalizes across multiple datasets,
* Handles rare diseases better,
* Is interpretable & clinically meaningful,
* And works in real-world conditions,

…it will stand out as *better* than the state-of-the-art.

Would you like me to create a **roadmap/plan** for your project (step-by-step tasks: dataset prep → model selection → training → validation → explainability → deployment)? This will help you implement these improvements in a structured way.

**Roadmap :**

Perfect ✅ — let’s build you a **practical, step-by-step roadmap** to make your skin disease detection project strong, novel, and ready for research-level presentation.

Here’s a **6-phase roadmap** you can follow:

**🗺️ Skin Disease Detection Model Roadmap**

**Phase 1: Problem Definition & Setup**

**Goal:** Define scope, gather resources, set up environment.

* ✅ **Define Task:** Multiclass classification (e.g., 7-class HAM10000) or binary (malignant vs benign)?
* ✅ **Identify Stakeholders:** Dermatologists/medical advisors if possible.
* ✅ **Environment Setup:**
  + Python (3.10+), PyTorch or TensorFlow/Keras
  + GPU runtime (Colab Pro / Kaggle / local GPU)
* 📦 **Tools to install:** torch, timm, albumentations, grad-cam, scikit-learn, pandas, matplotlib, wandb (for experiment tracking).

**Phase 2: Data Collection & Preprocessing**

**Goal:** Build a robust dataset pipeline.

* 📊 **Datasets:**
  + **Primary:** HAM10000 (10k images, 7 classes)
  + **Secondary (for validation):** ISIC 2018/2019, Derm7pt (for generalization testing)
* 🛠️ **Preprocessing Steps:**
  + Resize to fixed resolution (e.g., 224×224 or 384×384)
  + Remove image artifacts (rulers, color charts) if present.
  + Normalize pixel values to [0,1] or ImageNet mean/std.
  + Split: Train (70%) / Validation (15%) / Test (15%)
* 🎨 **Augmentation:** (via Albumentations)
  + Rotation, horizontal/vertical flip
  + Color jitter (brightness/contrast)
  + Random crop & resize
  + Gaussian noise (simulate different devices)
  + **Optional advanced:** Hair removal preprocessing (DullRazor algorithm)

**Phase 3: Baseline Model & Training**

**Goal:** Establish a strong baseline, then iterate.

* 🧠 **Model Choice:**
  + Start with **EfficientNet-B0** or **ResNet50** (transfer learning).
  + Use pretrained weights on ImageNet → fine-tune last layers.
* ⚙ **Training Setup:**
  + Loss: CrossEntropyLoss (or Focal Loss for imbalance)
  + Optimizer: AdamW, LR Scheduler: Cosine Annealing
  + Batch size: 32–64, Epochs: ~20–30
  + Metrics: Accuracy, F1-score, AUC (better for imbalanced data)
* 🧪 **Experiment Tracking:** Log results using W&B / TensorBoard.

**Phase 4: Model Improvement**

**Goal:** Go beyond baseline and outperform existing work.

* 🔁 **Iterate with stronger backbones:** EfficientNet-B3/B4, ConvNeXt, Vision Transformers.
* 🎯 **Add Attention:** SE blocks / CBAM for better feature focus.
* 🖼 **Multi-scale input:** Crop lesions + feed both full & zoomed-in images.
* ⚖ **Class imbalance:** Use weighted loss or oversample minority classes.
* 🤝 **Ensemble Models:** Average predictions from 2–3 architectures.

**Phase 5: Explainability & Validation**

**Goal:** Make the model trustworthy & robust.

* 🧐 **Explainability:**
  + Grad-CAM / Grad-CAM++ heatmaps to show focus area.
  + Overlay on lesion image for dermatologist validation.
* 🌍 **External Validation:**
  + Test on a dataset **not seen during training** (ISIC 2019).
  + Compare with state-of-the-art results from ISIC Challenge papers.
* 🩺 **Confidence Estimation:**
  + Calibrate probabilities (temperature scaling)
  + Show “Low Confidence” predictions for human review.

**Phase 6: Deployment & Documentation**

**Goal:** Make it usable and publishable.

* 🌐 **Deploy:**
  + Use **Streamlit or Gradio** for an interactive web app.
  + Optionally build a lightweight TFLite/ONNX model for mobile.
* 📜 **Documentation:**
  + Write a **literature review** citing existing works.
  + Compare your results (tables/graphs vs prior papers).
  + Highlight generalization, explainability, and real-world readiness.
* 🏆 **Publication Ready:**
  + Prepare report or paper with methodology, results, ablation studies, and conclusions.
  + Optionally submit to a student conference or ML competition.

**📊 Roadmap Visualization**

| **Phase** | **Key Output** |
| --- | --- |
| 1 | Defined problem + setup environment |
| 2 | Clean, augmented dataset + splits |
| 3 | Baseline EfficientNet model + metrics |
| 4 | Improved hybrid/ensemble model |
| 5 | Grad-CAM explanations + external validation results |
| 6 | Web demo + final report/paper |

Would you like me to create a **Gantt-style project timeline** (e.g., 8–10 weeks) with milestones for each phase? This could help you plan weekly goals and track progress systematically.

Perfect — here’s a **Gantt-style project timeline** you can follow for 8–10 weeks.  
I’ve broken it down week-by-week with **deliverables, milestones, and checkpoints** so you can track progress.

**📅 Skin Disease Detection Project – 10-Week Timeline**

| **Week** | **Phase / Focus** | **Tasks** | **Deliverables / Milestones** |
| --- | --- | --- | --- |
| **1** | **Phase 1 – Problem Definition & Setup** | Define project scope (classification, segmentation, or both)List diseases/classes to detectSet up development environment (Python, PyTorch/TensorFlow, GPU runtime) | ✅ Clear problem statement ✅ Git repo with environment setup (requirements.txt / conda env) |
| **2** | **Phase 2 – Data Collection** | Download HAM10000 & ISIC datasetsExplore dataset distribution (class imbalance, image size)Clean/rename files & create train/val/test splits | 📊 Dataset summary report (counts per class, sample images) ✅ Train/val/test split ready |
| **3** | **Phase 2 – Preprocessing & Augmentation** | Implement preprocessing pipeline: resizing, normalizationAdd augmentations (flip, rotate, crop, color jitter)Optional: implement hair-removal preprocessing | 🖼️ Augmentation examples (visualized) ✅ Preprocessing pipeline script ready |
| **4** | **Phase 3 – Baseline Model** | Load EfficientNet-B0 / ResNet50 pretrained modelFine-tune on datasetLog training metrics (loss, accuracy, F1-score) | 📈 Baseline results (accuracy, F1, confusion matrix) 📄 Short baseline report |
| **5** | **Phase 4 – Model Improvements (Round 1)** | Try a stronger backbone (EfficientNet-B3, ConvNeXt)Use Focal Loss / class weighting for imbalanceAdd early stopping + LR scheduler | 🚀 Improved model with better validation performance ✅ Comparison table vs baseline |
| **6** | **Phase 4 – Model Improvements (Round 2)** | Add attention mechanism (SE/CBAM)Experiment with multi-scale input (zoomed crops)Train & evaluate an ensemble of 2–3 models | 🏆 Best-performing model checkpoint 📊 Performance chart (accuracy, AUC, F1) |
| **7** | **Phase 5 – Explainability & Visualization** | Generate Grad-CAM/Grad-CAM++ heatmapsVisualize attention maps on test samplesAnalyze model failures & misclassifications | 🔎 Grad-CAM visual report 📝 Error analysis summary |
| **8** | **Phase 5 – External Validation** | Test best model on ISIC 2019 / Derm7ptCompare with ISIC leaderboard resultsCalibrate model confidence (temperature scaling) | 🌍 Generalization results ✅ Confidence calibration plot |
| **9** | **Phase 6 – Deployment Prep** | Build Streamlit / Gradio web appAllow users to upload images & view predictions + Grad-CAMOptional: export model to ONNX/TFLite for mobile | 💻 Working interactive demo ✅ Deployment documentation |
| **10** | **Phase 6 – Final Report & Presentation** | Write project report / thesis chapterInclude literature review, methodology, results, ablation studyPrepare slides & live demo | 📄 Final report (well-formatted) 🎤 Presentation slides & demo video |

**🔑 Tips for Execution**

* Track experiments using **Weights & Biases** or **TensorBoard** so you can compare models systematically.
* Keep each week’s results documented (accuracy tables, charts, sample predictions).
* Schedule **mini-review meetings** with your team/advisor every 2 weeks to stay aligned.
* Don’t wait until the end to think about deployment — start designing UI in week 6–7.

Would you like me to make a **visual Gantt chart (image)** of this timeline for you — so you can include it in your report or project proposal? It will look much cleaner and professional.